MATEMATIKA, 2013, Volume 29, Number 1c, 177-185

Estimation of Model Variance Functions in Survey Sampling using Historical Micro-data

Roberto Gismondi
ISTAT (Italian National Statistical Institute), Economic Statistics Department
Via Tuscolana 1788, Roma, Italy
e-mail: gismondi@istat.it

Abstract In this context, supposing a sampling survey framework and a model-based approach, the attention has been focused on the main features of the optimal prediction strategy of a given population mean, which implies estimation of some model parameters and functions, normally unknown. In particular, a wrong specification of the single unit model variances may lead to a serious loss of efficiency of estimates. For this reason, we have proposed some techniques for the estimation of model variances, which instead of being put equal to given *a priori* functions, can be estimated through historical data concerning past survey occasions. This approach is pragmatic and realistic, since quite always a time series of past observations is available, especially in a longitudinal survey context. Moreover, a simple post-stratification method has been proposed, in order to better define the models which can explain observed data. Finally, a comparative non parametric donor imputation procedure has been considered, which may be used separately or coupled with model assisted estimation. Usefulness of the techniques proposed has been tested through an empirical attempt, concerning the quarterly wholesale trade survey carried out by ISTAT (Italian National Statistical Institute) in the period 2005-2010. In this framework, the problem consists in minimizing magnitude of *revisions*, given by the differences between preliminary estimates (based on the sub-sample of *quick* respondents) and final estimates (which take into account *late* respondents as well). Main results show that model variances estimation through historical data leads to efficiency gains (lower average revisions) which cannot be neglected, and that model based prediction is normally more efficient than generalized regression estimation (which takes into account the sampling design randomness as well). Moreover, in many cases the *mixed procedure* (joint use of estimations of model unit variances through historical data, post-stratification and donor imputation) can improve precision of preliminary estimates even more.

Keywords Donor; Longitudinal Survey; Model; Non Response; Post-Stratification; Revision; Variance.

2010 Mathematics Subject Classification 62D05

1 Introduction

Along the last years, for sampling estimation purposes the recourse to a model based approach became more and more important, at least as an additional tool for increasing quality of sample estimates. On the other hand, the use of model based estimations in the context of official statistics is still scarce, in Italy as well as in the European Union context. That mostly depends on the objective risk due to the not correct knowledge of model parameters or functions which are necessary in order to implement estimates. From now on we will focus the attention towards the estimation of finite population parameters such as the mean or the total of a variable of interest y for a given finite population, and only *univariate* modelization will be taken into account.

According to a super-population approach, the optimal estimation strategy can be based on the minimization of the mean squared error (*MSE*) respect to the model that is supposed to explain observed data. The main risk is due to the need to identify the fittest model, taking into account that in a given domain of interest more than one model may occur [1]. Hedlin *et al.* [2] underlined the risk of additional bias due to a model miss-specification even when the asymptotically design unbiased *GREG* estimator is used. In particular, the issue of choosing a model has two main aspects: 1) its mathematical form; 2) the specification and estimation of parameters or model functions; quality improvements can result either from a more appropriate model or proper solutions as regards the second issue [3]. Recourse to a model based approach may be useful in presence of non response as well.

A *non respondent* is a generic unit whose data are not available at the estimation stage. The non response bias often depends on a model misspecification, for instance because respondents and not respondents follow different patterns [4]. Performances of traditional strategies for reducing non response bias are often poor, for instance because few auxiliary variables are available at the estimation stage. A late experience regarding employment data [5] showed that the non response bias may be not systematic, but may happen for some survey occasions and/or domains only. Moreover, many imputation techniques may not reduce bias enough to balance the increase of variance due to imputation [6].

In this context, we propose some techniques for improving the above mentioned aspect 2 (specification or model functions). After a resume of the basic known results concerning optimal prediction of a population mean under a linear model, we focus the attention on the need to use reliable estimates of model variance functions (formula (1) in section 2), whose correct specification is fundamental in order to achieve to not biased and efficient estimates. The estimation is driven by the availability of historical micro-data concerning the target variable observed in previous survey occasions, as it often happens in many real longitudinal survey contexts. Other criteria for improving estimates – which may be overlapped with the unit variance estimation techniques – have been also proposed (post-stratification and donor imputation). An empirical attempt has been proposed in section 3. Section 4 contains some perspective conclusions.

2 Materials and Methods

We indicate as i a generic population unit, $s$is the observed sample including n units, $\bar{s}$ is the not observed population, while the whole population, including N units, is $s \cup \bar{s}$. Current estimates refer to a period p of a year T. A period may be given by a month or a quarter; we suppose P periods in a year (P=12 or P=4). For each T and p, the purpose of the sample survey is the estimation of the population mean $\bar{y}_{Tp}$. For each unit we suppose the model:

$$y_{Tpi} = \beta_{Tp} x_{Tpi} + \varepsilon_{Tpi} \qquad \text{with:} \quad E(\varepsilon_{Tpi}) = 0 \qquad V(\varepsilon_{Tpi}) = \sigma^2_{Tp} v_{Tpi} \ \forall i \qquad Cov(\varepsilon_{Tpi}, \varepsilon_{Tpj}) = 0 \ \text{ if } i \neq j \tag{1}$$

where expected values E, variances V and covariances are referred to the model, x is an additional variable strongly correlated with y, the model variance function v_i is unknown, as well as β and σ^2. The model (1) is analogous to the one used in [5] in a sampling context for employment data where v=x for each unit i. If $\bar{y}_{Tps}$ is the sample mean, $f_{Tp} = n_{Tp} / N_{Tp}$, $x_{Tp\bar{s}}$ and $v_{Tp\bar{s}}$ are sums over units not in the sample, it is well known ([7], pp.385-390) that the optimal unbiased predictor (e.g. – defined as $\hat{\bar{y}}_{Tp}$ the predictor – such that $E(\hat{\bar{y}}_{Tp} - \bar{y}_{Tp}) = 0$ and which minimises the model MSE) of the not observed mean is:

$$\hat{\bar{y}}^*_{Tp} = f\,\bar{y}_{Tps} + (1 - f_{Tp})\bar{x}_{Tp\bar{s}}\beta^*_{Tp} \quad \text{where:} \quad \beta^*_{Tp} = \left(\sum_s \frac{x_{Tpi}\, y_{Tpi}}{v_{Tpi}}\right)\left(\sum_s \frac{x^2_{Tpi}}{v_{Tpi}}\right)^{-1} \tag{2}$$

where $\hat{\bar{y}}_{Tps}$ is the sample mean. The use of an estimator given by the sample mean is coherent with the homoschedastic model implied by $x_i=v_i=1$ for each i. When $v=1$, $v=x$ or $v=x^2$ one gets, respectively, that β is given by: 1) the *OLS* estimate when regression through the origin is used, 2) ratio between the y and x sample totals, 3) sample mean of ratios y/x. Two of the main risks underlying the model based prediction (2) are:

1) the model (1) may be not correctly specified. This risk should be evaluated carefully before the application of (1) and falls outside the methodological context herein discussed.
2) Even though the model (1) is correctly specified, it is not possible to use, in the prediction process, good estimates of the variables v_i.

A way to reduce risks due to a model-based strategy is the recourse to a mixed approach, based on both a model and a design driven inference. Under model (1), an alternative robust estimation strategy may be based on the generalized regression estimator ([7], p.399). Under simple random sampling, the *GREG* estimator is:

$$T_{Tp,GREG} = \bar{y}_{Tps} + \beta^*_{Tp}(\bar{x}_{Tp} - \bar{x}_{Tps}) . \tag{3}$$

The estimator T_{GREG} is unbiased respect to the model, it is asymptotically unbiased respect to the sampling design and its expected sampling variance respect to the model is the lowest in the class of design-unbiased predictors.

We propose a simple method for estimating each variance component v_i, which tries to use the real variability of historical data. From the original model (1) we have: $v_{Tpi} = V(y_{Tpi})/\sigma^2_{Tp} \approx \hat{V}(y_{Tpi})$, where the last term is an estimate of v_{Tpi} unless a constant term which in the second formula (2) disappears. We suppose to deal with a sample survey context for which a database of historical micro-data, derived from past survey occasions, is supposed to be available. The database includes micro-data referred to k consecutive years before T, so that, for each unit, it will contain kxP observations referred to the y variable object of interest. In this framework, for each unit i which is *respondent* as regards the period p in the year T, an estimate of the individual model variance will be given by:

$$V(\hat{\bar{y}}_{Tpi}) = \sum_{t=T-k}^{T} (y_{tpi} - \hat{\bar{y}}_{Tpi})^2/(k+1) \qquad \text{where:} \qquad \hat{\bar{y}}_{Tpi} = \sum_{t=T-k}^{T} y_{tpi}/(k+1) . \tag{4}$$

The estimation criterion (4) consists in the calculation, for each respondent unit, of an empirical longitudinal variance based on (k+1) addenda (k historical data and the actual observation at time T), where the second function in (4) is the empirical longitudinal mean. The main advantage derived from the use of observed historical variability of individual data is that it avoids any *a priori* exact - but dangerous - formulas for modelling unit variances. On the other hand, the use of (4) implies that a reliable estimate of the model variance functions v_i, which refer to a *certain period p of a given year T*, may be approximated by a *longitudinal* estimate, derived from a synthesis of the individual unit variability along time. If the unit i is *non respondent* as regards the period p in the current year T, an estimate of the individual model variance will be still based on (4), but using k observations only. Through (4) seasonality of estimates is saved, since each variance is estimated using past data referred to the *same* period p.

An *empirical* variance may be more reliable if a larger number of observations is used. A method for increasing the number of addenda consists in removing the seasonality constraint which has been implicitly supposed in (4). The counterbalance is given by the larger number of observations used for estimating variances. In symbols, we have:

$$V(\hat{\bar{y}}_{Ti}) = \sum_{t=T-k}^{T} \sum_{p=1}^{P} (y_{tpi} - \hat{\bar{y}}_{Ti})^2 / P(k+1) \qquad \text{where:} \quad \hat{\bar{y}}_{Ti} = \sum_{t=T-k}^{T} \sum_{p=1}^{P} y_{tpi} / P(k+1) \,. \tag{5}$$

If the criterion (5) is used, we adopt the *same* model unit variance estimates for any reference period *p*. A criterion similar to (5) is mandatory if we deal with a new survey: in this case the individual variances may be estimated starting from the period *p*=2, mixing in the variance formula the observations related to the periods *p*=1 and *p*=2.

A further criterion is still based on calculation of empirical variance estimates (4) or (5): we indicate with $\hat{V}(y_{Tpi})$ this estimate. Estimates (4) or (5) may contain outliers and, as a consequence, estimators (2) or (3) may be wrong; an alternative strategy consists in using (4) or (5) for the estimation of the parameters *a* and *b* of the first model in (6); after log-linearization, *OLS* estimates $\hat{a}$ and $\hat{b}$ may be used for calculation of the new variance estimates (second formula in (6)):

$$\hat{V}(y_{Tpi}) = a\, x_{Tpi}^{b} \qquad \rightarrow \qquad \hat{\hat{V}}(y_{Tpi}) = \hat{a}\, x_{Tpi}^{\hat{b}} \,. \tag{6}$$

According to observed data, it may be verified that within a given reference domain different models may exist, since individual expected values and/or variances could follow different patterns.

Post-stratification is often used in presence of non response problems, but is also a tool for testing the presence of different sub-populations in which model variances may be almost homoschedastic, In this case, the problem concerning estimation of model variances would disappear. In this context we propose a simple procedure based on a cluster analysis algorithm aimed at identifying *r* clusters. The sub-populations are supposed to be characterized by different levels of β and σ. As a consequence, for each period *p* the data matrix which can be used for clustering contains on the rows the single sample units and on the 2 columns the new variables z_{p1} and z_{p2} defined as follows:

$$z_{p1i} = \frac{1}{k+1} \sum_{t=T-k}^{T} \frac{y_{tpi}}{x_{tpi}} \approx \hat{\bar{\beta}}_{Tp(i)} \qquad z_{p2i} = \sqrt{\sum_{t=T-k}^{T} (y_{tpi} - \hat{\bar{y}}_{Tpi})^2 \Big/ \sum_{t=T-k}^{T} v_{tpi}} \approx \hat{\bar{\sigma}}_{Tp(i)} \tag{7}$$

where $\hat{\bar{y}}_{Tpi}$ has been defined in (4).

For each *p*, the first variable is an estimate of the "average slope" which characterises the *i-th* unit along the (*k*+1) survey occasions. The second variable is an estimate of the "average standard deviation" which characterises the same unit in the same time lag. Its formal structure derives from the general variance model formula: $\sigma_{Tp}^2 = V(y_{Tpi}) / v_{Tpi}$, from which an estimate of the average variance which characterises the *i-th* unit along the *k* past years is: $\hat{\bar{\sigma}}_{Tp(i)}^2 = \hat{V}(y_{tpi}) / \hat{\bar{v}}_{Tpi}$, where $\hat{V}(y_{Tpi})$ is given by the first formula (4) and $\hat{\bar{v}}_{Tpi} = \sum_{t=T-k}^{T} v_{tpi} / (k+1)$. The basic rationale is that each cluster should contain the units more similar to each other in terms of average slope and variability level through the recent past. After the identification of clusters, more estimation criteria can be applied inside each of them. When research is limited to 2 clusters, if one of the two clusters include only one unit, which is a non respondent, then the estimate of its *y*-level can be put equal to that obtained in the frame of the correspondent not post-stratified estimation strategy.

Dangerousness of model based estimation may suggest the recourse to non parametric estimation criteria. One of the most used, especially in the context of census surveys, is donor imputation. For each

$i \in \bar{s}$ a donor unit $j(i)$ is selected among the n available respondents' labels $j \in s$. The *nearest neighbor* method is based on the rule:

$$\hat{y}_{Tpid} = y_{Tpj(i)}(x_{Tpi} / x_{Tpj(i)}) \qquad \text{with:} \qquad D_{Tpj(i),i} = \left| x_{Tpj(i)} - x_{Tpi} \right| = Min_{j \in s}(D_{Tpi,j}) \qquad \text{for each } i \in \bar{s} \qquad (8)$$

where $D_{i,j}$ is a distance operator between the couple of units (i,j). The rationale of formula (8) is that, since units should follow a common response model, which however is not fully satisfactory explained through the expression (1), then the response provided by a respondent unit which is very similar to the non respondent one as regards the x variable (the donor) should be a good *proxy* of the unknown y value of the receiving unit. The additional use of the ratio between the two correspondent x values should correct for the residual distance between units $j(i)$ and i.

3 Results and Discussion

ISTAT (the Italian National Statistical Institute) is carrying out the "Wholesale trade and commission trade sector" (classification NACE Rev.2, division 46) since 2001. While provisional estimates – based on *quick* respondents – are released after 60 days from the end of the reference quarter, final indexes (including *late* respondents as well) are released after 180 days. The sampling survey is based on a stratified random sampling including about 7.500 units; the stratification considered in this context is based on four economic activities: 1) Wholesale on a fee or contract basis; 2) Agriculture raw materials and live animals; 3) Food, beverages, tobacco, household goods; 4) Non agriculture intermediate products, machinery, equipment, supplies, other products, and 3 employment classes (1-5 persons employed; 6-19; >19). Up to now, the implicit hypothesis maintained in the estimation approach is that late responses follow a *missing at random* (*MAR*) mechanism; that is the theoretical justification of the recourse to the current estimator given by the ordinary quick respondents sample mean, used both for provisional and final estimates.

In this context, the attention has been addressed towards the estimation of quarterly turnover means (instead of indexes). A longitudinal database has been built up for this purpose, including all and only the units belonging to the theoretical sample in each of the 6 years taken into account (from 2005 to 2010). The database contains the following variables: identification code, stratum code, quarterly turnover from first quarter 2005 until fourth quarter 2010, binary variable equal to 1 if a unit was respondent within 60 days and to 0 otherwise for each 2010 quarter, binary variable equal to 1 if a unit was a final respondent (within 180 days) and to 0 otherwise for each 2010 quarter. A crucial aspect concerned the choice of the auxiliary x variable. The empirical evidence (see also [8]) showed a very strong correlation between quarterly turnover and turnover related to the same quarter of the previous year (p-4), so that the final choice was $x_{Tp}=y_{T(p-4)}$.

The empirical exercise herein discussed is founded on the following rationale: we considered as the main object of estimation the *final* sample y mean (based on the N final respondents) and as estimator the prediction T based on the only n *quick* respondent units. In this way it was possible to calculate the real prediction error (revision) obtained using the various estimation strategies. The y means object of estimation concerned the four2010 quarters, the auxiliary x variables were turnover data referred to the four2009 quarters, while all the quarterly turnover data 2005-2010 have been used for implementing the empirical variance estimations described in section 2.

The average number of final respondents in the four quarters 2010 was 4,395. The number of final respondents ranged from 345 for the domain 2 (Agriculture raw materials and live animals) up to 1,957 for domain 3 (Food, beverages and household goods). The relative share of quick respondent units on final respondents ranged from 70.7% in the second quarter up to 86.2% in the third quarter. The

coefficient of variation of quarterly turnover was quite high on average, since it passed from 2.74 in the first quarter up to 2.88 in the third one.

Basically, two main estimators have been used and compared: the model based optimal predictor defined by (2) and the *GREG* estimator (3). They have been implemented in the following ways:

1) using a priori model variances, according to the common positions $v=1$, $v=x$ and $v=x^2$ for each unit.
2) Applying estimations of model unit variance based on historical data, on the basis of formula (4), putting $k=5$ (variances have been estimated using all the years from 2005 to 2009 for late respondents and from 2005 to 2010 for quick respondents), or $k=3$ (the years used ranged from 2007 to 2010 for quick respondents and from 2007 to 2009 for the late ones); moreover, estimates have been calculated using formula (5), which implies not seasonal variances (labeled as *Nseas* in the Table 1) and formula (6), which led to estimates labeled as *Model*.
3) On the basis of a *Pseudobest* strategy. For each of the strategies defined coupling a given estimator with a certain criterion for estimating model unit variances, it was possible to calculate the mean of absolute per cent errors of estimates (*MAPE*), evaluated on the *only quick respondent units*. For each domain and quarter the *Pseudo best* strategy was the one characterized by the lowest mean of errors. Of course this strategy minimizes *MAPE* based on quick respondents, but may not minimize *MAPE* of estimates which include late respondents as well. Risks of scarce efficiency will be high if response patterns of quick and late respondents are different.

Moreover, the previous estimation strategies have been implemented with or without the post-stratification method defined by (7), which has been developed putting $r=2$ and using the Ward algorithm. Each combination among kind of estimator ((2) or (3)), criterion for estimating model unit variances ((4), (5) or (6)) and use or not of post-stratification based on (7) identifies a specific *estimation strategy*.
Finally, the donor technique (8) has been used as well. However, in this case the method was applied in the following way: for each domain, quarter and strategy, the final estimator was given by the simple arithmetic mean between the estimator obtained on the basis of the given strategy and donor estimation. In this way we did not neglect the original model based approach neither when a non parametric estimation technique is introduced, but adopted a balanced mix between parametric and non parametric estimation techniques.

Each strategy has been applied separately in each sub-domain; sub-domains were 12 (4 main economic activities by 3 employment classes, already defined) or 24 if post-stratification was used as well. Then estimates for each domain and for the total wholesale trade sector have been obtained through weighted arithmetic mean of sub-domain estimates, were weights derive from structural business statistics. Goodness of estimates has been evaluated on the basis of *MAPE.*

The main results have been resumed in the Table 1. For each strategy (columns), *MAPE* has been reported for the average of the 4 economic activities concerned and the total wholesale trade (rows). As a matter of fact, the strategy given by the sample mean of quick respondents – used as reference benchmark – led to the worst estimates for any domain; *MAPE* was very high: it was equal to 8.34 for the average of 4 economic activities and to 9.36 for total wholesale trade.

If we consider the optimal model based prediction, an important outcome is that all the strategies based on estimation of model variances based on time series (which will be defined as *new strategies* from now on) led to average *MAPEs* lower than those obtained using the common position $v=1$, $v=x$ or $v=x^2$ (*basic strategies* from now on). The best strategies were *Nseas* – with *MAPE* equal to 1.19 – and *5 years* – with *MAPE* equal to 1.22. Among the basic model based predictions, the position $v=x$ performed as the best (*MAPE*=1.51). The *Pseudo best* strategy (last column) guaranteed good but not optimal results (*MAPE*=1.41).If we consider estimates of the Total wholesale trade (row "Total"), strategies *5 years* and *Nseas* are still the best, with quite similar *MAPE*s (0.68 and 0.70 respectively) and the performance of the *Pseudo best* strategy is very good (*MAPE*=0.78). Also in this case, the *new strategies* performed better than the *basic* strategies, with the only exception of *3 years*, whose *MAPE* is larger than those obtained

using the position v=1 (1.06) and v=x (1.07).On the whole, gains in precision of estimates derived from the use of historical data for estimation of model variances are clear and encouraging.

On average, the joint recourse to post-stratification and donor imputation led to better results than the use of one of the two single criteria separately, even though important efficiency gains derived from post-stratification have been obtained as regards the Total wholesale trade (but not as regards the average of 4 economic activities). Overall, as concerns the average of 4 quarters, efficiency gains with respect to results obtained without post-stratification and/or donor imputation have been obtained with the positions v=x^2 (*MAPE* decreases from 1.87 to 1.74) and using the *Pseudo best* strategy (*MAPE* decreases from 1.41 to 1.37). A quite better performance has been obtained for the Total wholesale trade: there are efficiency gains using all the various criteria (with the exceptions of *5 years* and *Nseas*, whose *MAPE*s remain steady), even though gains are quite more relevant with the *basic strategies* rather than with the *new* ones. We can conclude that, when a model based optimal prediction is used, the additional support provided by post-stratification and donor imputation is particularly important just when the huge longitudinal variability of the y-variable may be a serious obstacle to the correct estimation of unit variances.

Table 1 MAPEs obtained with the use of the optimal model based prediction, GREG and various options

	Sample	***A priori* model variances**			**Model variances based on**				***Pseudo***
Estimation domains	***mean***	***v*=1**	***v*=*x***	***v*=x^2**	***5***	***3 years***	***Nseas***	***Model***	***Best***
Optimal model based prediction									
Average (1-4)	8.24	1.76	1.51	1.87	1.22	1.42	1.19	1.50	1.41
Total	9.36	1.06	1.07	1.55	0.68	1.17	0.70	1.00	0.78
Optimal model based prediction and post-stratification									
Average (1-4)	8.24	1.82	2.09	2.22	1.88	1.89	1.70	2.05	1.78
Total	9.36	0.79	0.60	0.99	0.95	0.64	0.60	0.68	0.82
Optimal model based prediction and donor imputation									
Average (1-4)	8.24	2.52	2.12	2.12	1.91	2.08	1.90	2.23	2.04
Total	9.36	1.53	1.22	1.26	1.01	1.13	1.08	1.32	1.16
Optimal model based prediction with post-stratification and donor imputation									
Average (1-4)	8.24	1.87	1.66	1.74	1.36	1.51	1.34	1.67	1.37
Total	9.36	0.88	0.94	1.44	0.68	1.05	0.70	0.95	0.74
GREG									
Average (1-4)	8.24	2.52	2.12	2.12	1.91	2.08	1.90	2.23	2.04
Total	9.36	1.53	1.22	1.26	1.01	1.13	1.08	1.32	1.16
GREG and post-stratification									
Average (1-4)	8.24	1.73	1.66	1.62	1.43	1.51	1.41	1.66	1.43
Total	9.36	0.80	0.94	1.18	0.74	0.95	0.74	0.92	0.75
GREG and donor imputation									
Average (1-4)	8.24	2.71	2.67	2.68	2.45	2.54	2.46	2.68	2.51
Total	9.36	1.24	0.95	0.69	1.18	0.89	1.17	0.99	1.19
GREG with post-stratification and donor imputation									
Average (1-4)	8.24	1.54	1.51	1.50	1.31	1.42	1.30	1.50	1.44
Total	9.36	0.94	1.07	1.30	0.79	1.10	0.82	1.03	0.84

"Average (1-4)" is the mean of *MAPE*s concerning domains: 1) Wholesale on a fee or contract basis; 2) Agriculture raw materials and live animals; 3) Food, beverages and household goods; 4) Other products. "Total" refers to "Total wholesale trade".

The use of the *GREG* estimator did not lead to efficiency gains. Indeed, on average of 4 economic activities its performance was always significantly worse than that of the optimal model based predictor, for any unit variance estimation criteria adopted. However, for the Total wholesale trade there were two efficiency gains: the former concerns v=x^2 and is significant (*MAPE* decreases from 1.55 of the optimal

model based predictor to 1.26), while the latter concerns *3 years* and is small (*MAPE* decreases from 1.17 of the optimal model based predictor to 1.13).

On the other hand, the joint use of post-stratification and donor imputation produced significant efficiency gains for *GREG*. *MAPE* of the *Pseudo best* strategy passed from 2.04 to 1.44 as regards the average of 4 economic activities and from 1.16 to 0.84 as regards the Total wholesale trade. The most relevant result is that, as concerns the average of 4 economic activities, post-stratification and donor imputation performed better when coupled with *GREG* rather than with the optimal model based prediction. For instance, this evidence characterized all cases when new criteria for estimating model variances are applied: *5 years* (*MAPE* decreases from 1.36 to 1.31), *3 years* (from 1.51 to 1.42), *Nseas* (from 1.34 to 1.30) and *Model* (from 1.67 to 1.50). Efficiency gains have been also obtained when basic strategies for model unit variance estimation are used. On the other hand, as regards the Total wholesale trade, post-stratification and donor imputation still perform better when coupled with the optimal model based predictor rather than *GREG*.

4 Conclusion

In order to apply model based estimation in current surveys, it is needed to have reliable estimates of model parameters and functions. The availability of historical micro-data concerning the same target survey can be useful for implementing simple techniques aimed at obtaining estimates of the model unit variance functions, which must be always specified in each non homoschedastic model. These techniques are founded on the idea to approximate each model unit variance function with the empirical unit longitudinal variance calculated through the historical database. Moreover, the implicit dangerousness of a model based approach may suggest to search for a better model detection (post-stratification), and/or to use a non parametric estimation method – as donor imputation – as well.

In this context, different estimation techniques based on historical data have been proposed, depending on the number of observations to be included in calculations, on the importance of seasonal effects and on the opportunity to use an estimation based both on the empirical variance criterion and on log-linear estimation. An empirical attempt, based on the quarterly sample survey on wholesale trade carried out by ISTAT, confirmed the usefulness of the new approaches for estimating model unit variances, in comparison with other basic *a priori* assumptions. As regards the best estimation technique, the main conclusions are in favor of the model based optimal prediction when post-stratification and donor imputation are not used, and in favor of *GREG* when they are jointly used.

Further research is needed in order to face two main issues, with the goal of enforcing results herein obtained:

- additional comparative applications are needed, based on other longitudinal surveys and other response rates. Efficiency of techniques compared should be investigated unless the presence of non (or late) responses, even though the true values of the population mean should be always available, in order to calculate real *MAPE*s.
- Other *y* variables should be taken into account. Turnover represents a not easy task, since other variables – as the number of persons employed – may be more steady. However, discontinuity along time may be larger for investments and changes of stocks, or when the target variable is a binary variable (number of job vacancies).

References

[1] Ibrahim, J.G., Zhu, H. and Tang, N. Model Selection Criteria for Missing-Data Problems Using the EM Algorithm. *Journal of the American Statistical Association*. 2008.103(484): 1648-1658.

[2] Hedlin, D., Falvey, H., Chambers, R. and Kocic, P. Does the Model Matter for *GREG* Estimation? A Business Survey Example. *Journal of Official Statistics*. 2001. 17(4): 527-544.
[3] Lehtonen, R., Särndal, C.E. and Veijanen, A. The Effect of Model Choice in Estimation for Domains, Including Small Domains. *Survey Methodology*. 2003. 29(1): 33-44.
[4] Slud, E.V. and Bailey, L. Evaluation and Selection of Models for Attrition Nonresponse Adjustment. *Journal of Official Statistics*. 2010. 26(1): 127-143.
[5] Copeland, K.R. and Valliant, R. Imputing for Late Reporting in the U.S. Current Employment Statistics Survey. *Journal of Official Statistics.* 2007. 23(1): 69-90.
[6] Watson, N. and Starick, R. Evaluation of Alternative Income Imputation Methods for a Longitudinal Survey. *Journal of Official Statistics*. 2011. 27(4): 693-715.
[7] Cicchitelli, G., Herzel, A. and Montanari, G.E. *Il campionamento statistico*. Bologna: Il Mulino. 1992.
[8] Gismondi, R. Reducing Revisions in Short-term Business Surveys. *Statistica*. Anno LXVIII. 2008. 1: 85-116.