Estimation procedure and inference for component totals of the economic aggregates in the "Frame SBS"[1]

Paolo Righi[2]

Abstract

Recently the Italian National Institute of Statistics - Istat - implemented a Business frame where several variables are collected from administrative registers. Nevertheless, these variables do not cover all statistical interests and some variables are collected only by the Small and Medium Enterprise survey – SME survey. The paper deals with the estimation of totals of variables strictly observed in Istat SME survey and proposes an estimation procedure, based on the projection estimator, exploiting the variables of the Business frame and coherent with respect to the totals of the variables in the frame. The result is an integrated output in the Business frame and a flexible tool useful for other statistical purposes. Inferential properties are shown theoretically and empirically and conditions to obtain unbiased estimates are pointed out.

Keywords: Administrative data sources, projection estimator, design based inference.

1. Introduction

Most of the new Istat Business *frame* variables (Luzi e Monducci, 2016) come from the several Italian administrative data sources. They cover only partially the business economic information demand. Nevertheless, other variables and the respective parameters such as totals or means are required by EU Structural Business Statistics (SBS) Regulation. In particular, they are fundamental for implementing econometric models analyzing trend and the performance of the economic system. Usually such variables represent the *components* of economic aggregates which are known from the archives or are imputed in the frame by previous steps (see Di Zio *et al*., 2015).

The only direct informative source of these components is essentially the Small and Medium Enterprises (SME) sampling survey conducted by Istat. In the SME survey the calibration estimator (Deville and Särndal, 1992) is used. However, the large amount of auxiliary information, now available, is inefficiently exploited by this estimator. Furthermore, the output of the estimator is not suitable for the frame purposes. In the paper we propose an estimation process, exploiting the auxiliary variables in an enhanced way, for the totals of these components. The new estimator takes into account some appealing requirements: the sum of the estimated totals for the elementary components belonging to a given economic aggregate must be coherent to the (estimated) total of the economic

[1] The views expressed in this paper are solely those of the author and do not involve the responsibility of Istat.

[2] Istat, e-mail: parighi@istat.it

aggregate at domain level according to the current SBS Regulations; the output should be a flexible statistical tool and it can be used for other aims. In particular, the Istat National Account (NA) sector bases its procedures on the frame, so the coherence of the estimates of the components must be fulfilled for the NA domains that are generally highly detailed.

To obtain these objectives, the *projection estimator* (Kim and Rao, 2012) has been used. The method imputes or *projects* the component values of the not sampled enterprises in the SME survey by using an estimated regression models. The final estimates are achieved as the sum of the projected values by the models.

There are some advantages in using the projection estimator. At first glance the micro level estimates (projected values) seem the most appealing feature of the estimator. Nevertheless, such feature has to be used carefully because it hides a dangerous drawback of producing biased estimates at certain level of detail (section 2.1). The most relevant properties involve the inferential process. The projection estimator takes into account the randomization process of the SME sampling design and the inference is performed by a model assisted approach. That greatly simplifies the computation of the precision of the estimates, especially when a large scale population (about 4.4 millions of enterprises) has to be investigated. Model assisted approach is, commonly, used in the national statistical office and the variance estimation of the projection estimator can be found in classical textbooks. Moreover, the approach guarantees unbiased and robust estimates at least at certain domain level (see section 2.1) without an overwhelming model diagnostic required when a model based approach is taken into account.

Finally, the projection estimator is a more flexible tool compared to the generalized regression estimator (Särndal *et al.*, 1992), approximating the calibration estimators. The regression (and calibration) estimator considers a unique set of covariates in the regression model; the projection estimator varies the set of covariates in the regression model when the variables of interest change. That means each component is projected by a specific statistical model and that allows the improving precision of the estimates.

These conditions justify the choice to identify the projection estimator as a tool to complete the Business frame.

The outline of the paper is as follows: section 2 is devoted to the description of the projection estimator, highlighting the theoretical aspects and the bias issue. Section 3 describes the practical implementation of the estimator. Since the SME sample is affected by unit non response, the weight adjustment process for unit nonresponse is shown. The projection estimator has been implemented using the adjusted sampling weight. Section 4 gives an approximate estimate of the sampling errors. Section 5 presents brief conclusions.

2. Projection estimators

The projection estimator was introduced long ago in the sampling literature, but recently has had considerable attention (Hidiroglou, 2001; Merkourios, 2004; 2010) and the paper by Kim and Rao (2012) well formalizes the fundamental properties. Schenker and Raghunathan (2007) reported several applications using a model-based inference. Unlike Kim and Rao proposed a model assisted framework that is robust against failure of the working model used to generate the synthetic or projected values.

The estimator arises to deal with a nonnested two-phase sampling design. This design

involves two independent surveys from the same target population U consisting of N elements. A large sample s_1 from survey 1 collects information only on a vector of variable $\mathbf{x} = (x_1,...,x_q,...x_Q)'$ and a much smaller sample s_2 from survey 2 provides information on both y and $\mathbf{x}$, being y the variable of interest. It is assumed that the observed variables $\mathbf{x}$ are comparable. The concept of comparability refers to the classical test theory in which two kinds of measurement errors are distinguished (Bakker, 2012): validity and reliability. According to McCall (2001), reliability refers to whether the measurement procedures assign the same value to a characteristic each time it is measured under essentially the same circumstances. Unreliable measurement leads to random error. Validity refers to how accurately the values assigned in the measurement procedures reflect the actual conceptual variable measured. Invalid measurement leads to systematic error or bias in estimates (McCall, 2001). In the following we make the approximation of the absence of two kinds of measurement errors.

The main aim of the estimator is in creating a single synthetic dataset of proxy values $\tilde{y}_k$ (k=1, …, N) for the unobserved y_k values in survey 1 and then using the proxy data together with the associated survey weights, w_{1k} of survey 1 to produce projection estimates of the population and domain (or subpopulation) totals of y. Since the estimator creates an imputed dataset associated with the sample s_1, the method is classified as mass imputation technique too.

We focus on the estimator of the totals $Y_d = \sum_{k \in U} y_k \delta_k(d)$, where $\delta_k(d)$ is the domain membership indicator variable. The total for the overall population is a specific case obtained setting $\delta_k(d)$=1 always. The sub-population total is obtained setting $\delta_k(d)$=1 if unit $k \in U_d$ and $\delta_k(d)$=0 otherwise being U_d the dth domain (d=1, …, D).

Projection estimator is assisted by a superpopulation working model. Let a general formulation of the working model be $E(y_k \mid \mathbf{x}_k) = f(\mathbf{x}_k, \boldsymbol{\beta}) = \mu_k$, with $Var(y_k \mid \mathbf{x}_k) = \sigma^2 a(\mu_k)$ for some known function $a(\mu_k)$ and that $Cov(y_k, y_j \mid \mathbf{x}_k, \mathbf{x}_j) = 0$ for $k \neq j$. For a continuous variable y as the case of the components to be projected in the frame the linear model is a suitable choice. The working model is fitted by relating y to $\mathbf{x}$ using the data $\{(y_k, \mathbf{x}_k) : k \in s_2\}$ and $\tilde{y}_k = f(\mathbf{x}_k, \hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ is obtained as a solution to

$$\sum_{k \in s_2} w_{2k} [(\partial \mu_k / \partial \boldsymbol{\beta}) / a(\mu_k)](y_k - \mu_k) = 0 ,$$

according to the estimation function theory (Godambe and Thompson, 1986). We point out that the ordinary and weighted least square methods for linear model belong to this class of parameter estimators.

Finally, the projection estimator at domain level is given by

$$\hat{Y}_{d,p} = \sum_{k \in s_1} \tilde{y}_k \delta_k(d) w_{1k} . \qquad (2.1)$$

In our estimation context we assume the SME survey as the second survey while the

first large survey is the business register covering the entire population, being w_{1k}=1. In the SME survey and in the business frame the $\mathbf{x}$ variables are the economic aggregates and some other auxiliary variables such as number of employed persons and economic activity. The $\mathbf{x}$ variables of both the data sources are comparable being the systematic errors and the unreliability reduced (Luzi e Monducci, 2016). The estimator (2.1) becomes

$$\hat{Y}_{d,p} = \sum_U \tilde{y}_k \delta_k(d) \qquad (2.2)$$

As far the overall population total is concerned the projection estimator is given by $\hat{Y}_p = \sum_U \tilde{y}_k$.

2.1 Bias and variance

The working model introduced in section 2 is domain independent. However, there are some advantages to consider the domain if we want produce unbiased estimates.

Usually the estimator (2.2) produces biased estimates being an approximate expression of the bias given by

$$B(\hat{Y}_{d,p}) \cong \sum_{k \in U} \delta_k(d)[y_k - f(\mathbf{x}_k, \boldsymbol{\beta}_0)], \qquad (2.3)$$

in which $\boldsymbol{\beta}_0$ is the probability limit of $\hat{\boldsymbol{\beta}}$ with respect to the second sampling design. An estimate of the domain bias is given by $\hat{Y}_{d,bc} = \sum_{k \in s_2} \delta_k(d) w_{2k}(y_k - \tilde{y}_k)$, so a bias-corrected version domain estimator is

$$\hat{Y}_{d,p,bc} = \hat{Y}_{d,p} + \hat{Y}_{d,bc} . \qquad (2.4)$$

Unlike the projection estimator (2.2) the bias corrected estimator requires the use of the data and of the survey weights from the second survey and the issue could be unattractive.

Nevertheless, there are some conditions in which (a) $\hat{Y}_{d,bc}$=0 or (b) the bias of the estimator (2.3) is asymptotically negligible with respect to the domain total Y_d .

The condition $\hat{Y}_{d,bc}$=0 is achieved when the $\mathbf{x}_k$ vector include the $\delta_k(d)$ value. That means the domain intercept has to be included in the linear model underling the projection estimator. When a heteroscedastic linear model is used, with $Var(y_k \mid \mathbf{x}_k) = \sigma^2 x_{qk}$ then the variable $\delta_k(d) x_{qk}$ must be included in the regression line in order to obtain $\hat{Y}_{d,bc}$=0.

It is worthwhile to note that the bias-corrected estimator has internal consistency property: if condition (a) is fulfilled for a domain U_d the estimates when summed over the sub-domains defining a partition of U_d agrees with $\hat{Y}_{d,p}$.

As far condition (b) is concerned the asymptotic bias of the projection domain estimator relative to the domain total is given by

$$RB(\hat{Y}_{d,p}) = -\frac{N\,Cov(\delta_k(d), r_k)}{Y_d}, \qquad (2.5)$$

where $Cov(\delta_k(d), r_k)$ is the population covariance of $\delta_k(d)$ and $r_k = y_k - f(\mathbf{x}_k, \boldsymbol{\beta}_0)$. It follows from (2.5) that $RB(\hat{Y}_{d,p})$ is negligible if $\delta_k(d)$ is approximately unrelated to r_k. Roughly speaking, such condition is verified when in the scatter plot of the *d*th domain points is fairly distributed over and under the regression line. The expression (2.5) is equivalent to the formula (13) proposed by Kim and Rao (see Appendix 1). The (2.5) highlights that for large Y_d the relative bias becomes relatively small.

As far the variance is concerned, in the standard nonnested two-phase sampling design estimator (2.1) an approximate expression, when the condition (a) or (b) holds, is

$$Var(\hat{Y}_{d,p}) \cong Var_1[\sum_{k \in s_1} \delta_k(d) w_{1k} f(\mathbf{x}_k, \boldsymbol{\beta}_0)] + Var_2[\sum_{k \in s_2} \delta_k(d) w_{2k} (y_k - f(\mathbf{x}_k, \boldsymbol{\beta}_0))], \qquad (2.6)$$

where $Var_1[.]$ and $Var_2[.]$ are the design variances respectively of the first and the second sampling design. In our survey context is quite dissimilar from the usual one if we treats the first survey as census. In this case $Var_1[.]$ disappears and the variance of the projection estimator becomes

$$Var(\hat{Y}_{d,p}) \cong Var_2[\sum_{k \in s_2} \delta_k(d) w_{2k} (y_k - f(\mathbf{x}_k, \boldsymbol{\beta}_0))], \qquad (2.7)$$

which is the standard formula used for the generalized regression estimator. By the consequence we may use the standard variance estimator of the generalized regression estimator (Särndal *et al.*, 1992). The assumption is that the economic aggregate **x** values are really observed. Ignoring the imputation process implemented for some units (Di Zio *et al.*, 2015) the expression (2.7) is a downward variance approximation. The goodness of the subsequent inference will depend on the performances of the imputation step and the rate of the imputed values.

Introducing the imputation uncertainty the variance expression becomes more complex (Appendix 2) and it is not dealt with in the application.

3. Estimates of the economic component totals: application of the projection estimator

The procedure is based on the sample of respondents of the SME survey. Bias conditions and variance of the projection estimator have been taken into account for setting the regression models. There is a trade-off between bias and precision; models defined including the domain intercept at highly detailed domain level allows to compute unbiased detailed estimates, but variance estimation could increase. So we cannot use too specific regression models. The estimation procedure considered the coverage of about 33,600 respondents of the whole population (SME survey, year 2011). The analysis led to consider

models for domains defined according to the Nace Rev. 2 three digit economic activity by the size class of employed persons (0-5, 6-19, 20-99).

We started with about 600 domains (and regression models) with generally a minimum number of sampled units equal to 25 and an average number of about 45 of sampled units. In some cases we obtained smaller sample size domains but with a high sample rate (Table 3.1). Finally, it was necessary to collapse some economic activities / or classes of employed persons to gather an enough number of respondents for estimating the models, obtaining 583 domains.

Table 3.1 - Rules for not collapsing the domain

(If) Number of respondents	(than) Sample rate (respondents/population size) must be
1-2	1.00
4-5	0.50
6-8	0.10
9-14	0.02

Under this level of detail the estimates could be significantly biased according to the condition (2.5) and a simple solution it should be use the estimator (2.4), guaranteeing the internal consistency although the variance problem still remains. Otherwise, the small area estimation approach (Rao, 2003) could be used for more reliable estimates, but the procedure could be complex if internal consistency must be satisfied.

SME survey is affected by unit nonresponse. So we used the Response Homogeneity Group technique (Oh and Scheuren, 1983) to adjust the sampling weights for unit nonresponse. The sampling weights are inflated by the inverse of non-response rate measured at RHG level, being the sample size of 2011 SME survey of about 97,000 enterprises. After studying the best way to deal with the nonresponse the RHGs coincided with the domains of regression models. In particular, the logistics model and different nonresponse classes for nonresponse adjustment has been compared. The results have been not significantly different from the ones using the domains of the projection estimator. On the other hand implementing the logistic model for estimating the response probability can be cumbersome if the process has to be performed in each survey occasion.

The regression models assisting the *projection estimator* have been defined taking into account the space of the possible projection values. For all the components the constraint is to obtain non negative projected values but the components of the change economic aggregates. So, for the former type of variables the heteroscedastic ratio models have been used, where each component has as covariate the economic aggregate to which it belongs to. We point out that with this model the sum of the components of a given aggregate is equal to economic aggregate at enterprises level. The regression model for the components of the change economic aggregate uses standard heteroscedastic model where the heteroscedastic term is the square root of the number of employees.

4. Sampling errors of the projection estimator: some evidences

The efficiency of the projection estimator has been compared with the one of the

calibration estimator currently used in the SME survey. The analysis of the results has to envision two issues: (i) the variance of the projection estimator (computed on 33,600 respondents) does not take into account the previous imputation step on the economic aggregates; (ii) the variances of the calibration estimator are computed on the respondents and the integrated non respondents of the SME survey (73,200 enterprises) according to the procedure described by Casciano *et al.*, (2012). We point out that the component values of the integrated non respondents are imputed but the estimator treats them as if they were observed and the accuracy of the estimates will be generally overstated (Kalton and Kasprzyk, 1986; Righi *et al*., 2014). We remark a fundamental different role of the imputed values in the two sampling contexts. In the projection estimator, the imputed variables are the auxiliary variables, while in the current estimation strategy they are the interest variables. That means: the true projection estimator variance will be larger than the variance measured in the analysis; bias is introduced in the calibration estimator and the true mean square error will be larger than the one observed in the analysis.

Section 2.1 introduces the complexity for tackling the point (i). To deal with the point (ii) it should be necessary to known the imputation procedure, making the comparison too burdensome. Therefore, we consider the results as general evidences of the two estimator performances, underling when the imputation step affects the final evaluation.

Figures 4.1, 4.2 and 4.3 depict the Coefficient of Variations (CVs) of the projection and the calibration estimators for the totals for the entire target population. For sake of brevity, only some of the most important component variables are shown: *income from sales and services* (*turnover*), *purchases of services*, *purchases of goods*, *use of third party assets* and *other operating charges*.

Generally, the projection estimator outperforms the current estimation procedure. The results on the purchases of services components are more controversial. Especially for the components with small amount (and large CV), sometime the current procedure shows lower CV than the projection estimator. The calibration estimator outclasses the projection estimator also the component *C12905* of the other operating charges.

Figure 4.1 – CV (%) of the components (label from SME questionnaire) belong to *income from sales and services* realized by the generalized regression and projection estimator.

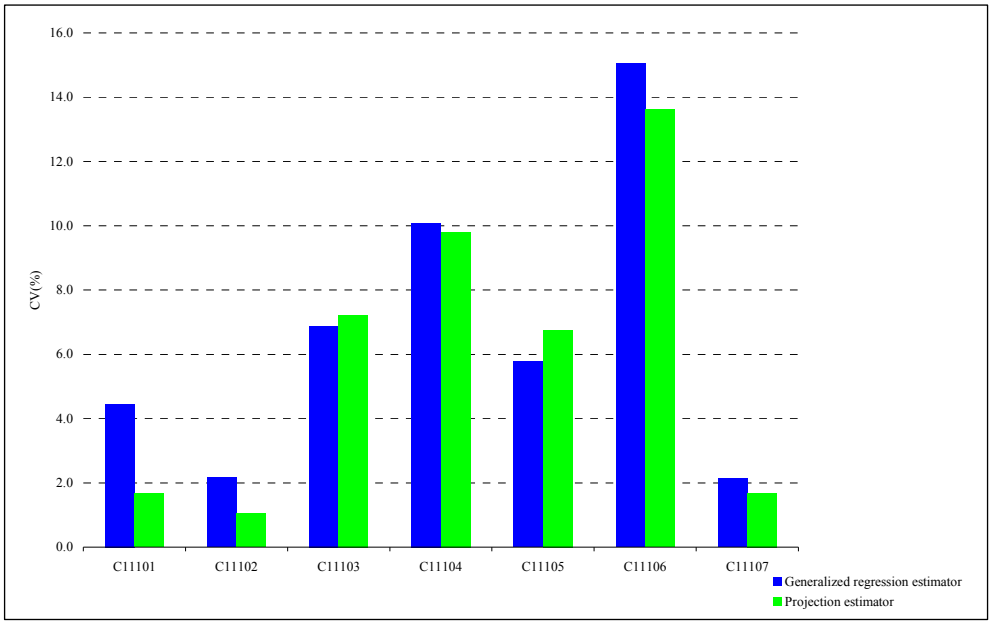


Figure 4.2 – CV (%) of the components (label from SME questionnaire) belong to *purchases of services* realized by the generalized regression and projection estimator.

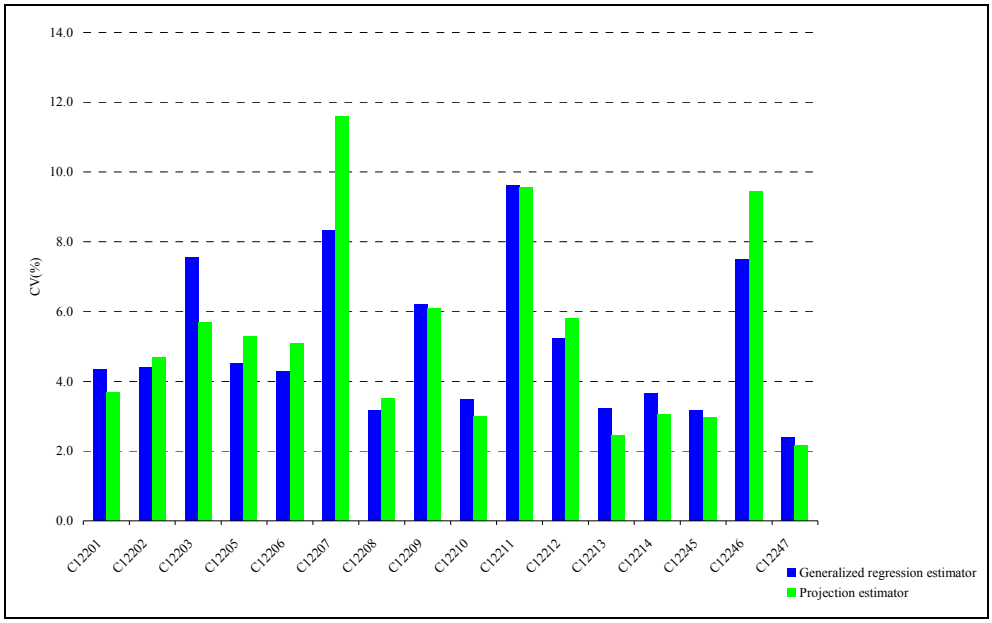
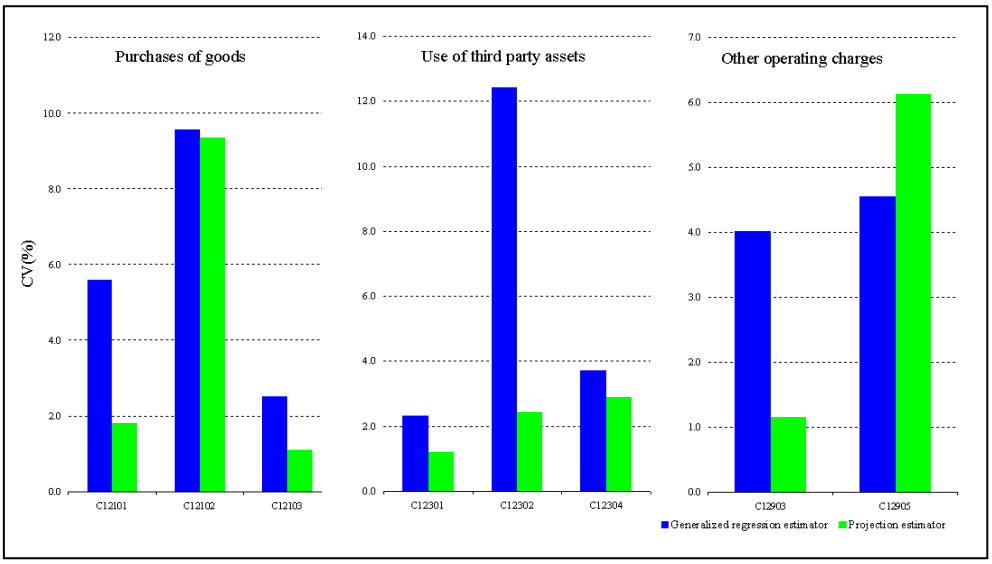


Figure 4.3 – CV (%) of the components (label from SME questionnaire) belong to *purchases of goods*, *use of third part assets* and *other operating charges* realized by the generalized regression and projection estimator.



The CV computed for the specific estimators at overall population level represents the average performance of such estimators a domain level.

Nonetheless, to get a really insight into the performances of the two estimators, we studied the CV distributions at domain level as well. The analysis reverses the relationships between the two estimators and the calibration estimator looks like better than the projection estimators. In particular, the former one produces lower CVs for totals relatively small (figure 4.4, 4.5 and 4.6). Figures 4.7 shows the median of the projection estimates of each component observed in the distribution of the domain estimates. When the median is quite small the CV distribution of the projection estimator is worse than the calibration estimator distribution. This evidence probably depends on the different sample sizes used since, for rare phenomena (or small amounts), the number of units have a greater impact on the precision on the estimates so ignoring the imputation step in the current estimation strategy the bias could be prevalent. On the other hand, the projection estimator shows his weakness for the small area estimation as usual for a direct estimator.

Figure 4.4 – Distribution of the CV (%) of the 583 domains for the components (label from SME questionnaire) belong to *income from sales and services* realized by the projection estimator (Projection) and the generalized regression estimator (GREG).

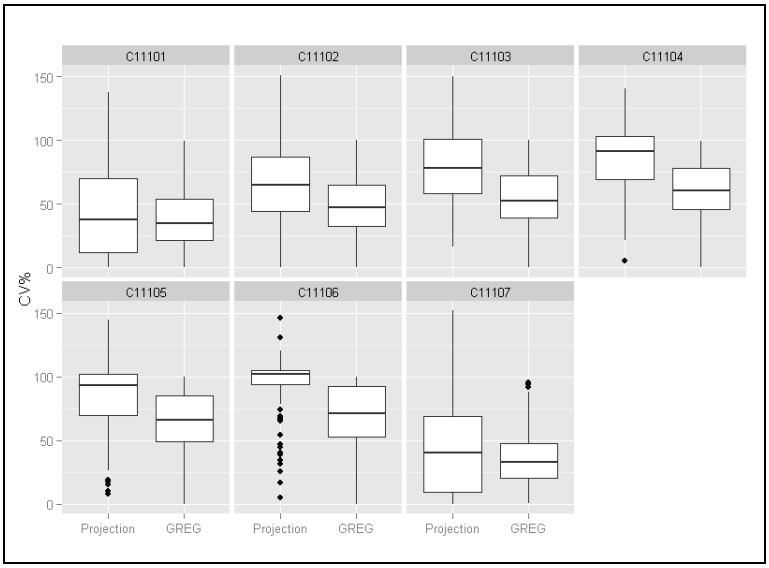


Figure 4.5 – Distribution of the CV (%) of the 583 domains for the components (label from SME questionnaire) belong to *purchases of services* realized by the projection estimator (Projection) and the generalized regression estimator (GREG).

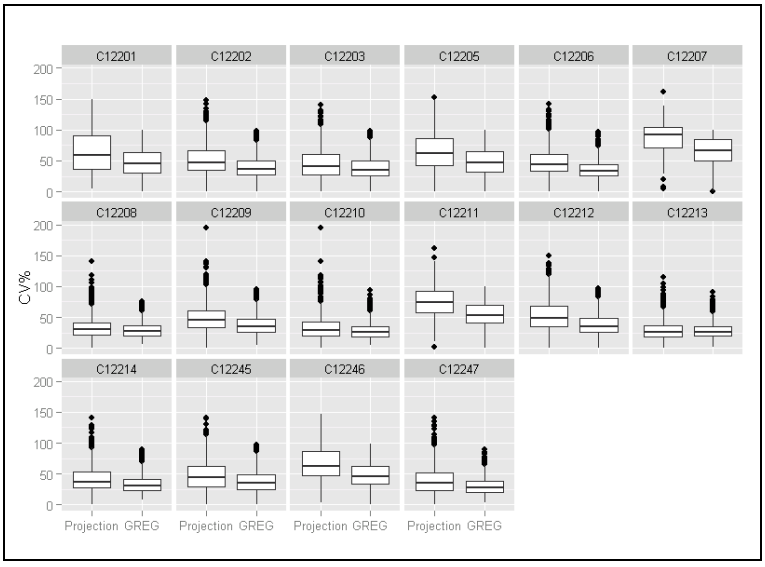


Figure 4.6 – Distribution of the CV (%) of the 583 domains for the components (label from SME questionnaire) belong to purchases of goods (upper left), use of third party assets (upper right)and other operating charges (lower left) realized by the projection estimator (Projection) and the generalized regression estimator (GREG)

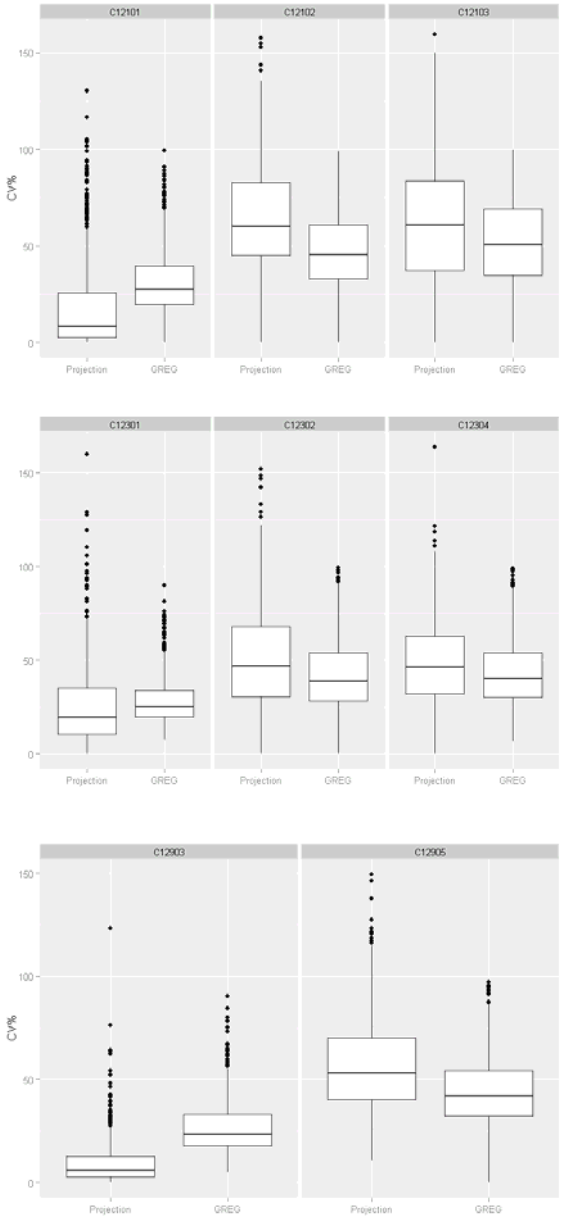
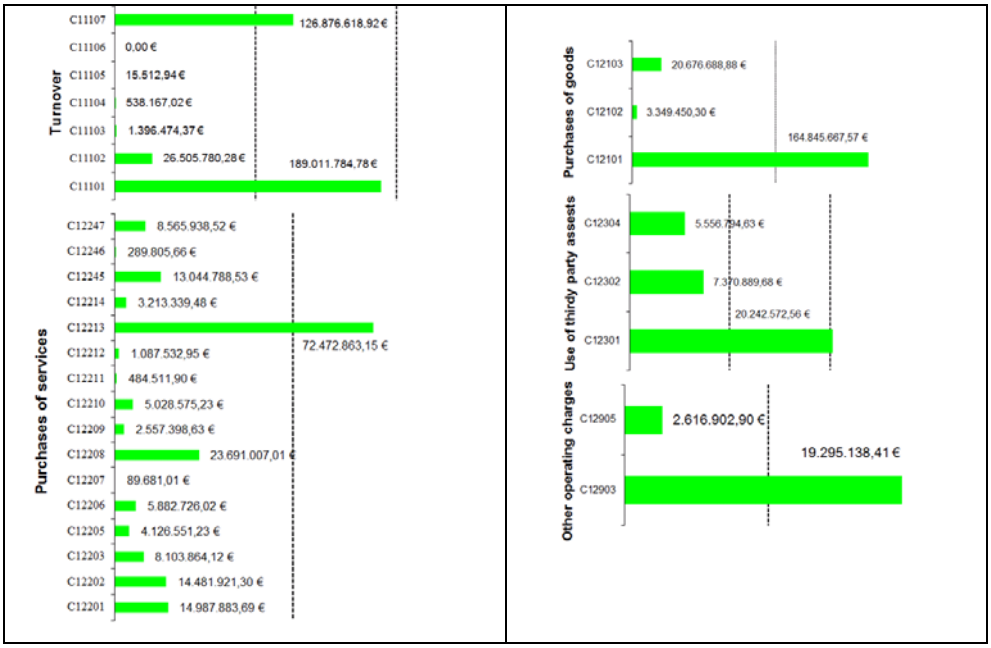


Figure 4.7 - Median values of the domain estimate distribution obtained by the projection estimator



5. Conclusions

The administrative data sources such as Balance sheets, Sector Studies, Tax returns, etc., although offer a large amount of economic variables are not exhaustive of the business information demand. The paper shows an estimation procedure, based on the projection estimator, to complete the set of estimates of the new Italian Business frame (Luzi e Monducci, 2016). The choice of using the projection estimator comes from a mix of operative conditions, theoretical properties and applicative opportunities. The estimation process is involved in a general context in which large data set and highly detailed domains are deemed. So an automatized and easy to implement method is quite appealing. The proposed process meets requirements and offers a well-founded inferential framework in which the sampling errors and bias are simple to compute and internal consistency is always satisfied. The outcome of the process is the input of other statistical processes. In particular, the Istat National Account (NA) sector bases its procedures on the frame and the imputation carried out by the projection estimator give interesting applicative opportunities. Anyway the projected values are not the true values and the inference must be carried out carefully at certain level of detail. In this case some tricks can be used. Otherwise, other estimation approaches, such as small area estimators, must be applied with the risk to complicate the sampling strategy. The projection estimator has been implemented from 2010 data onwards. The paper focuses on the precision of the estimates of the totals and a comparison with the current Structural Business Statistics estimates based on the SME survey is performed (year 2011). The SME survey uses the calibration estimator based on

the sample of respondent and the integrated non respondent. So the estimator uses about the double number of units with respect to the projection estimator that considers the sample respondents only. The findings have to be assessed taking into account that a part of the variance of the two estimators is ignored: in fact, the imputed values of the integrated respondents (for the current procedure) and the imputed economic aggregate variables of the frame (for the projection estimator) are treated as if they were observed.

As main results, the proposed technique for non small areas outperforms the current estimation strategy, because the auxiliary information of the new business frame are powerful predictors of the interest variables, underlying that the new estimation strategy will enhance the quality of the business statistics.

When in a given domain either or both the phenomenon is rare or the sample size is small (small area estimation problem) the comparison seems to be highly affected by the number of units used in the two estimators. In this case the performances of the current procedure is favored because a double sample size is used. As general indication, the result highlights a possible mean square error underestimation of the current procedure. As far the projection estimator is concerned, there are large CVs in many domains (Nace Rev. 2 3 digit by size class) even though the worst values should be only for the residual domains because for the overall population totals the CVs (considered as an average of the CV domain estimates) are quite low. The evidence recommends of using very carefully the estimates at high level of detail.

For these residual domains it should be better to use suitable estimators as small area estimators. But in this case other issues should be opened: integrate model assisted and model based estimates; know the domain types involved in the procedure (they are the domains of SBS Regulation, the domains of NA sector or types of domains that are not possible to foresee before processing the data); define a time spending process relate to the dimension of the data set, number of estimates and the estimation procedure itself.

Eventually, the proposed estimation procedure does not take properly into account the model uncertainty due to the imputation step of the economic aggregates implemented for some enterprises of the business register. The matter should be dealt with in the future for achieving a correct inferential analysis using data from the frame.

Appendix 1.

Proof of formula (2.5). Let us consider the expression (13) proposed by Kim and Rao (2012)

$$RB(\hat{Y}_{d,p}) = -\frac{Cov(\delta_k(d), r_k)}{\bar{\delta}(d)\bar{Y}_d},$$

where $\bar{\delta}(d) = (1/N)\sum_U \delta_k(d)$ and $\bar{Y}_d = (1/N_d)\sum_U y_k \delta_k(d)$.

Then

$$RB(\hat{Y}_{d,p}) = -\frac{N\,Cov(\delta_k(d), r_k)}{N\bar{\delta}(d)\bar{Y}_d} = -\frac{N\,Cov(\delta_k(d), r_k)}{[\sum_U \delta_k(d)][(1/N_d)\sum_U y_k\delta_k(d)]}$$
$$= -\frac{N\,Cov(\delta_k(d), r_k)}{\sum_U y_k\delta_k(d)} = -\frac{N\,Cov(\delta_k(d), r_k)}{Y_d}.$$

Appendix 2.

For obtaining a reduced downward approximate expression of the variance let us consider the working model used for imputing the missing values of the economic aggregate in the frame (Di Zio *et al*., 2015). We reformulate the first addendum of formula (2.6) as

$$Var_1[\sum_{U_O} f(\mathbf{x}_k, \boldsymbol{\beta}_0) + \sum_{U_M} f(\tilde{\mathbf{x}}_k, \boldsymbol{\beta}_0)], \qquad (A.1)$$

where the sample s_1 is replaced by $U = U_O \cup U_M$ with U_O and U_M respectively the population with observed and missing values and $\tilde{\mathbf{x}}_k$ the vector of imputed covariates in the frame. For sake of brevity we suppose a common pattern of missingness among the variables x_{qk} (q=1, …, Q). The Var_1 operator reflects the design variance so we need to introduce the model uncertainty of the previous imputation step.

Assume that the imputation of the x_q is ruled by the model $E_M(x_{qk} \mid \mathbf{z}_k) = g(\mathbf{z}_k, \boldsymbol{\gamma}) = \tilde{x}_{qk}$, with $Var_M(u_{qk} \mid \mathbf{z}_k) = \psi^2 b(\mathbf{z}_k)$, being u_{qk} the residual term, for some known function $b(.)$ and that $Cov_M(u_{qk}, u_{qj} \mid \mathbf{z}_k, \mathbf{z}_j) = 0$ for $k \neq j$, where the operators $E_M(.)$, $Var_M(.)$ and $Cov_M(.)$ are referred to the M imputation model. Since the expected values are equal to the true values, the model expectation is unbiased. Instead of the design variance we jointly consider the model and design variance. The model variance influences only the first addendum of the expression (2.6). Then we replace the expression (A.1) with

$$E_p E_M[\sum_{U_O} f(\mathbf{x}_k, \boldsymbol{\beta}_0) + \sum_{U_M} f(\tilde{\mathbf{x}}_k, \boldsymbol{\beta}_0) - \sum_U f(\mathbf{x}_k, \boldsymbol{\beta}_0)]^2, \qquad (A.2)$$

where $E_p(.)$ is the design expectation. Nevertheless the operator $E_p(.)$ disappears because we have a census and it can be shown that the (A.2) is equal to $\sum_{U_M} E_M\left\{f[g(u_{qk}, \boldsymbol{\gamma})]\right\}^2$. In case of $f(.)$ and $g(.)$ are linear function the model variance becomes $\sum_{U_M} \psi^2 b(\mathbf{z}_k)\boldsymbol{\beta}_0$.

References

Bakker B. F. M. 2012. "Estimating the validity of administrative variables". *Statistica Neerlandica*, 66: 8-17.

Casciano M. C., V. De Giorgi, F. Oropallo, G. Siesto. 2012. "Estimation of Structural Business Statistics for Small Firms by Using Administrative Data". *Rivista di Statistica Ufficiale,* n. 2-3/2012.

Deville, J.-C., C.-E. Särndal. 1992. "Calibration estimators in survey sampling". *Journal of the American Statistical Association*. 87: 376-382.

Di Zio M., U. Guarnera, R. Varriale. 2015. "The estimation of the main variables of the economic account of small and medium enterprises based on administrative sources". *Rivista di Statistica Ufficiale*, n. 1/2016.

Godambe V., Thompson M. 1986. "Parameters of superpopulation and survey population: their relationship and estimation". *International Statistical. Review*, 54: 127-38.

Hidiroglou M. 2001. "Double sampling". *Survey Methodology*, 27: 143-54.

Luzi O., R. Monducci. 2016. "The new statistical register "Frame SBS": overview and perspectives". *Rivista di Statistica Ufficiale*, pp. 5-14.

Kalton G., D. Kasprzyk. 1986. "The Treatment of Missing Survey Data". *Survey Methodology*, 12: 1-16.

Kim J. K., J.N.K. Rao. 2012. "Combining data from two independent surveys: a model-assisted approach". *Biometrika*, 99: 85-100.

McCall R. B. 2001. *Fundamental statistics for behavioural sciences*. Wadsworth, Belmont.

Merkouris T. 2004. "Combining independent regression estimators from multiple surveys". *Journal American Statistical Association*, 99: 1131-9.

Merkouris T. 2010. "Combining information from multiple surveys by using regression for efficient small domain estimation". *Journal of Royal Statistical Society B*, 72: 27-48.

Särndal, C. E., B. Swensson and J.H. Wretman. 1992. Model-assisted Survey Sampling. New York: Springer.

Oh H.L., F.J. Scheuren. 1983. Weighting adjustment for unit nonresponse, in: W.G. Madow, I. Olkin and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys*, Vol. 2. New York: Academic Press: 143-184.

Rao J.N.K. 2003. *Small Area Estimation*. Wiley, New York.

Righi P., Falorsi S., Fasulo A. 2014. "A modified Delete a Group Jackknife variance estimator under random hot deck imputation in business surveys". In F. Mecatti F., P. L.Conti and M. G. Ranalli (eds), *Contributions to Sampling Statistics*. Springer: 219-233.

Schenker N., T. Raghunathan. 2007. "Combining information from multiple surveys to enhance estimation of measures of health". *Statist. Med.*, 26: 1802-11.