ITACOSM 2019 - Survey and Data Science

Florence

5^{th} June 2019

The marginal impact of auxiliary totals in calibration

Claudio Ceccarelli ISTAT <clceccar@istat.it>

Alessio Guandalini ISTAT <alessio.guandalini@istat.it>

Introduction

Calibrated estimator (Deville & Särndal, 1992)
is widely used for deriving survey estimates

Main reasons (Särndal, 2007):

- increasing accuracy
- achieve consistent estimates
- just a set of sampling weights

Introduction

Calibrated estimator (Deville & Särndal, 1992)
is widely used for deriving survey estimates

Main reasons (Särndal, 2007):

- increasing accuracy
- achieve consistent estimates
    - very important for NSIs
    - means for promoting credibility in published statistics
    - better provision of administrative data and registers
- just a set of sampling weights

Introduction

Achieve consistent estimates

- ✓ higher level of consistency
- ✗ problem of convergence for the constrained optimization problem
- ✗ decreasing on the accuracy of estimates (*backfired effect*)

Introduction

Achieve consistent estimates

✓ higher level of consistency

✗ problem of convergence for the constrained optimization problem

✗ decreasing on the accuracy of estimates (*backfired effect*)

Impact of auxiliary totals

- best fitting

Introduction

Achieve consistent estimates

- ✓ higher level of consistency
- ✗ problem of convergence for the constrained optimization problem
- ✗ decreasing on the accuracy of estimates (*backfired effect*)

Impact of auxiliary totals

- best fitting
- Shapley decomposition (Shapley, 1953)

Calibrated estimator

"[...] a strong correlation between the auxiliary variables and the study variable means that the weights that perform well for the auxiliary variables also should perform well for the study variables".

Deville Särndal (1992) p. 376.

$$\hat{Y}_{CAL} = \sum_{k \in s} y_k \; w_k = \sum_{k_s} y_k \; d_k \; \gamma_k$$

$$\begin{cases} \min_{w_k} \; \sum_{k \in s} G\left(w_k, d_k\right)/q_k \\ \sum_{k \in s} w_k \; \boldsymbol{x}_k = \boldsymbol{X} \end{cases}$$

$$\text{var}\left(\hat{Y}_{CAL}\right) = \sum_{k \in s} \sum_{\ell \neq k} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \left(e_k w_k\right)\left(e_\ell w_\ell\right)$$

Shapley decomposition

based on the well-known concept of **Shapley value** (Φ) in cooperative game theory (Shapley, 1953)

$$\Phi\left(i,\nu\right)=\sum_{S\subseteq N}\frac{\left(|N|-|S|\right)!\left(|S|-1\right)!}{|N|!}\left(\nu\left(S\right)-\nu\left(S\right)/\left\{i\right\}\right)$$

re-assign the profits generated by a coalition of N players in proportion to the contribution that each player has made to the coalition itself.

Shapley decomposition

based on the well-known concept of **Shapley value** (Φ) in cooperative game theory (Shapley, 1953)

$$\Phi\left(i,\nu\right)=\sum_{S\subseteq N}\frac{\left(\left|N\right|-\left|S\right|\right)!\left(\left|S\right|-1\right)!}{\left|N\right|!}\left(\nu\left(S\right)-\nu\left(S\right)/\left\{i\right\}\right)$$

re-assign the profits generated by a coalition of N players in proportion to the contribution that each player has made to the coalition itself.

- application fields
  - regressors in a linear model (Israeli, 2007)
  - inequality measure decomposition (Shorrocks, 2013)

Shapley decomposition

based on the well-known concept of **Shapley value** (Φ) in cooperative game theory (Shapley, 1953)

$$\Phi(i, \nu) = \sum_{S \subseteq N} \frac{(|N| - |S|)!\,(|S| - 1)!}{|N|!} \left(\nu(S) - \nu(S) / \{i\}\right)$$

re-assign the profits generated by a coalition of N players in proportion to the contribution that each player has made to the coalition itself.

- application fields
  - regressors in a linear model (Israeli, 2007)
  - inequality measure decomposition (Shorrocks, 2013)
  - **auxiliary totals in calibration**

Example

linear regression model: $Y \sim X + Z + W \quad r^2_{Y/XZW} = 0.755$

Example

linear regression model: $Y \sim X + Z + W$ $\quad r^2_{Y/XZW} = 0.755$

	model	r^2
1.	Y	0.000
2.	$Y \sim W$	0.352
3.	$Y \sim Z$	0.122
4.	$Y \sim X$	0.503
5.	$Y \sim Z + W$	0.379
6.	$Y \sim X + W$	0.808
7.	$Y \sim X + Z$	0.755
8.	$Y \sim X + Z + W$	0.755

Example

linear regression model: $Y \sim X + Z + W \qquad r^2_{Y/XZW} = 0.755$

	model	r^2
1.	Y	0.000
2.	$Y \sim W$	0.352
3.	$Y \sim Z$	0.122
4.	$Y \sim X$	0.503
5.	$Y \sim Z + W$	0.379
6.	$Y \sim X + W$	0.808
7.	$Y \sim X + Z$	0.755
8.	$Y \sim X + Z + W$	0.755

Shapley value

$$\begin{array}{lcrr} \Phi\left(X, r^2\right) & = & 0.4745 & (62.8\%) \\ \Phi\left(Z, r^2\right) & = & 0.0695 & (9.2\%) \\ \Phi\left(W, r^2\right) & = & 0.2110 & (27.9\%) \\ \hline & & 0.7550 & (100.0\%) \end{array}$$

Application on calibration context

- players $\rightarrow$ auxiliary totals
- characteristic function $\nu(\cdot)$
  - change in estimate w.r.t. HT
  - change in variance in terms of $cv\%$ w.r.t. HT

Application on calibration context

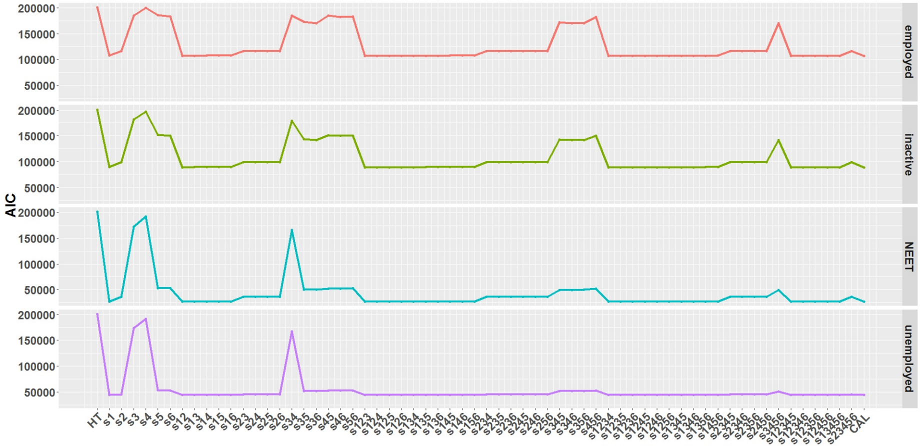
- players $\rightarrow$ auxiliary totals
- characteristic function $\nu(\cdot)$
  - change in estimate w.r.t. HT
  - change in variance in terms of $cv\%$ w.r.t. HT

re-assign the change generated by a **calibration system of N auxiliary totals (groups of auxiliary totals)** in proportion to the contribution that each auxiliary totals (group of auxiliary totals) has made to the coalition itself in terms of **change in the estimates** and **change in variance**.

Application on Italian LFS 2013-2016

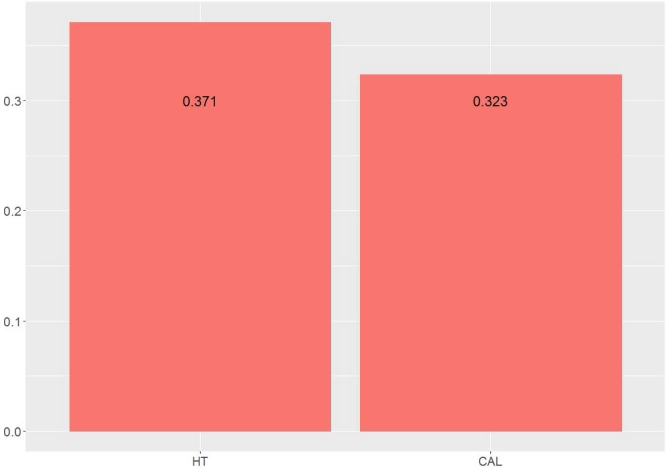
- players $\rightarrow$ auxiliary totals
  - X1 population by region, sex and 14 age classes [28x21]
  - X2 population by province, sex and 5 age classes [120x21]
  - X3 population by metropolitan city and 5 age classes [30x21]
  - X4 population of foreigners (M, F, UE, no-UE) by region [4x21]
  - X5 population by region and rotational group [4x21]
  - X6 population by region, sex and month in the quarter [6x21]
- characteristic function $\nu(\cdot)$
  - statistics (Y)
    - employment rate
    - unemployment rate
    - inactivity rate
    - NEET rate
  - domains
    - national (NUTS-0)
    - regional (NUTS-2)

Best fitting criterion



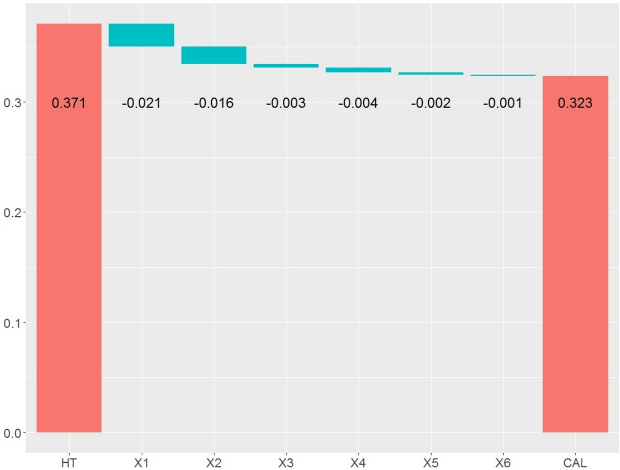
Change in variance ($cv\%$)

Employment rate - Italy, Q1-2013



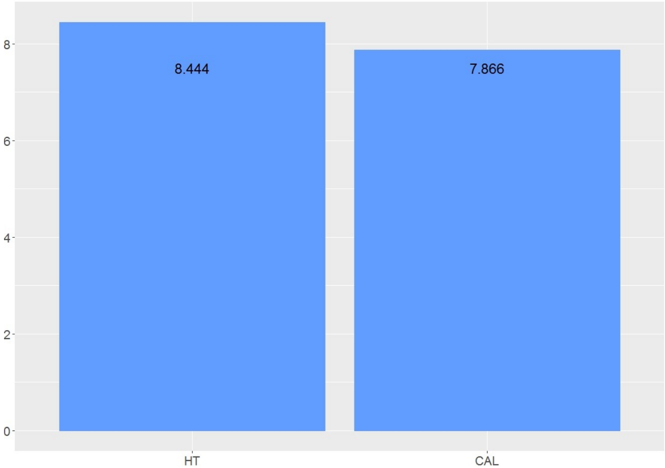
Change in variance ($cv\%$)

Employment rate - Italy, Q1-2013



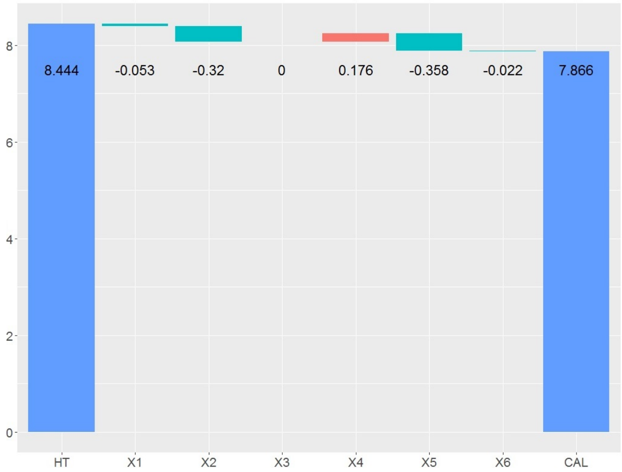
Change in variance ($cv\%$)

Unemployment rate - Friuli Venezia Giulia, Q1-2013

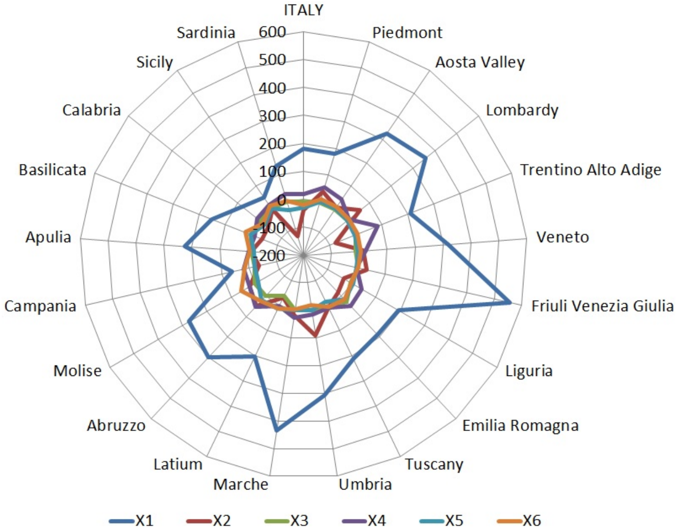


Change in variance ($cv\%$)

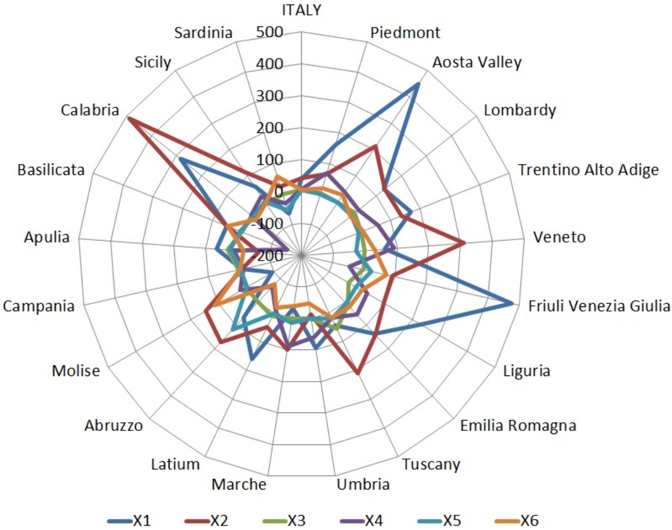
Unemployment rate - Friuli Venezia Giulia, Q1-2013



Marginal impact on estimates, Q1-2015

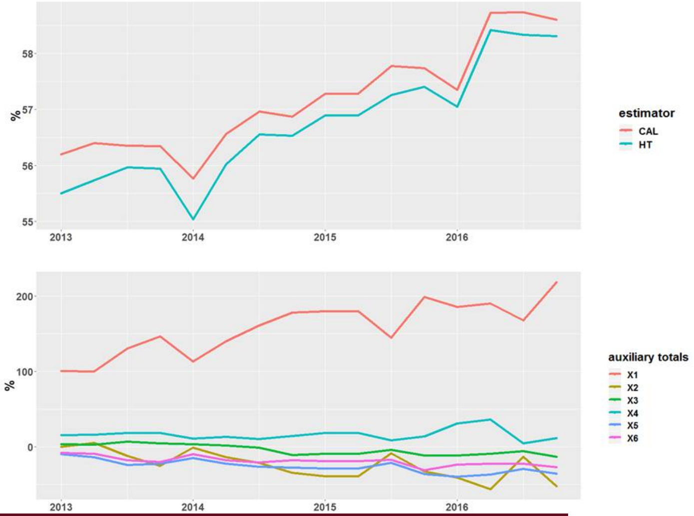


Marginal impact on variance (*cv%*), Q1-2015



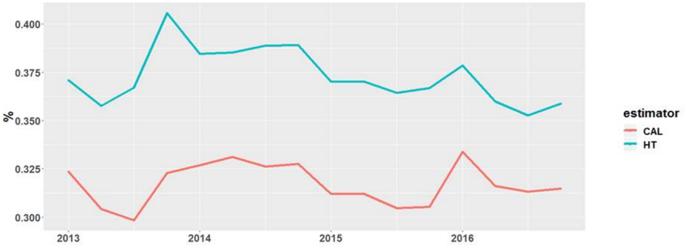
Marginal impact during the time (*estimate*)

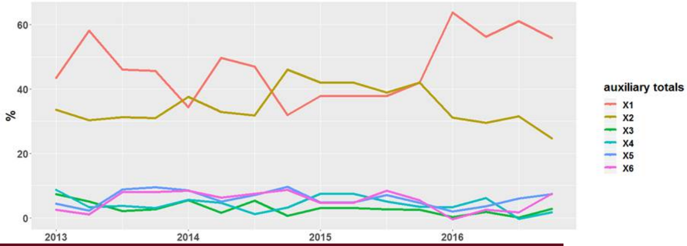
Employment rate - Italy



Marginal impact during the time ($cv\%$)

Employment rate - Italy





Conclusion

- Shapley decomposition applied to the calibration context can help to better understand the role in the estimation of auxiliary totals
- the use of auxiliary totals must be analyzed both in estimates and variance perspective
- the impact of auxiliary totals change w.r.t statistics and domains

References

- Deville J.-C., Särndal C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.
- Israeli O. (2007). A Shapley based decomposition of the R-square of a linear regression. *Journal of Econominc Inequality*, 5:199-212.
- Istat (2006). La rilevazione sulle Forze di Lavoro: contenuti, metodologie, organizzazione. *Metodi e Norme*, 32.
- Särndal C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33:99–119.
- Shorrocks A.F. (2013). Decomposition procedures for distributional analysis: a unified framework based on the Shapley Value. *Journal of Economic Inequality*, 11:99-126.