The Effect of Business Demography on Index Numbers Calculation

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

Abstract

In the business short-term statistics context – in particular, in the monthly index number calculation framework – business demography is a methodological issue not well focused and managed up to now. The reference population existing in the index base year changes month by month, because of business demography: some units die, some modify their economic features and new ones are born. Two basic methodological problems should be tackled: 1) the current monthly (or quarterly) sample, selected from the target population in the base year, provides information on deaths, but no information on births, so that the infra-year calculation of index numbers cannot be based on proper estimates of the new population size. 2) New units born along the reference year cannot be observed before the first month (or quarter) of the next year: if the longitudinal profile of new units is different with respect to the already existing ones, current estimates may be biased. In this work, we formalize the index number formula in a sampling survey context and compare the right index number which takes into account business demography with the one which ignores that, A criterion for correcting the sampling weights in order to overcome the business demography effects is also proposed. Moreover, we formalize and comment on a rotating sampling technique which includes new units in January of each new year, for taking into account the longitudinal profile of new (born) units. The usefulness of the previous proposals has been tested through a simulation study, based on real Italian quarterly wholesale trade turnover data referring to the period 2005-2010.

Keywords: Bias, Sampling, Short-term statistics, Stratification, Weighting

1. Introduction

Since 1998, in the European Union (EU) the Short-Term Statistics Regulation is ruling the calculation of index numbers concerning production, turnover and employment in almost the economic sectors, from industry to market services. Quality issues are ensured by specific features of consistency and exhaustiveness. However, there are some degrees of freedom in the calculation procedures which may be used by National Statistical Institutes. A relevant problem concerns the effects of business demography, which is not clearly stated in the Regulation and whose statistical treatment may be different depending on the country. The reference population existing in the base year – actually 2010 – changes month by month, because of business demography: some units die (definitive closure, economic activity changes), some modify their economic features (main economic activity, size, location) and others are born. However, almost in every EU country the business register is updated on a yearly basis and with about one year delay. As a consequence, two basic methodological problems should be tackled, resumed as follows: 1) the current monthly or quarterly sample, selected from the target population in the base year, provides information on deaths, but no information on births, so that the infra-year calculation of index numbers cannot be based on proper estimates of the new population size. Along the reference year, sampling weights will refer to the previous year population, which does not take into account balance between births and deaths. This issue may be very dangerous especially as regards young and very changing economic activities. 2) New

units born along the reference year cannot be observed before the first observation period of the new year, e.g. as soon as the new business register lets the possibility to update the sample composition: if the longitudinal profile of new units is different with respect to the already existing units, current estimates based on units existing in the base year only may be biased.

With these premises, in this work we first introduce the formalization of the right index number, e.g. the one which takes properly into account the effects of business demography on sampling weights. The potential bias of estimates due to the use of wrong sampling weights follows straightforwardly. Moreover, the formal difference between the indexes which include or not include the different longitudinal profile of "new" units is also quantified. A simple criterion for correcting the sampling weights in order to overcome the business demography effects is proposed, as well as a simple sampling design for reducing the potential bias due to exclusion of new units from the sample. The usefulness of the previous proposals has been tested through an empirical attempt, based on real Italian quarterly wholesale trade turnover data (2005-2010).

2. Sampling weights adjustment for business demography

The main goal of each monthly index (the extension to quarters is trivial) is to estimate the ratio between the population *y*-total and the average monthly population total in the base year. We suppose that in the base year T a unique sample s_T is observed along all the 12 months (the extension to quarterly estimates may be obtained easily). N_T is the population size known at the beginning of each year, while n_{Tm} is the sample size in the month m of year T (which can be supposed unchanged for any m unless wave non-response occurs). In the year $(T+k) - k=1,2..$ – in each month we suppose that the questionnaire asks for the *y*-variable amount in the month m of year $(T+k)$ and in the same month of the previous year $(T+k-1)$. This strategy is quite used in Italy (monthly retail trade survey, quarterly wholesale trade survey, both finalized at estimating turnover indexes). The possibility to calculate the ratios $m/(m-12)$ for each sample unit allows the check of *tendency* changes (comparison with the same month of the previous year) and the managing of wave non-response and the inclusion of "born" units into the sample (since for these units no past observation was observed in the past). The main goal consists in using chained indexes, based on the monthly calculation of the ratio between the *y*-total sample estimates related to the months m and $(m-12)$, using at *both* times the *same* sample units.

Under simple random sampling (*srs*), each sampling weight in the base year is $w_{iTm} = N_t / n_{tm}$, in the year $(T+1)$ the sampling weight for estimating the *y*-total at time $(T+1,m)$ is $w_{i,T+1,m} = N_{T+1} / n_{T+1,m}$, while the sampling weight for estimating the *y*-total at time (T,m) is $\hat{w}_{iTm} = N_T / n_{T+1,m}$, because we remind that the two last estimates are both based on the number of sampling units $n_{T+1,m}$ observed in the month m (Dufour *et al*., 2001). The extension to whatever year $(T+k)$ follows straightforwardly. The index calculation process is founded on the calculation of chained indexes obtained, for each month m, starting from the base year index and the recursive series of products with the yearly ratio-types $m/(m-12)$ related to the following years. If i is a sampling unit – selected in the base year – and $I_{Tm/T}$ is the index of month m, year T with base the year T, the index calculation scheme from the base year T up to a generic further year $(T+k)$ can be summarized in the table below. We have the following indexes referred to the month m: in the first row concerning the base year T, in the second row the year $(T+1)$ and in the third row concerning a generic year $(T+k)$. The (*srs*) design is also described in the second column. All formulas hold for any $m=1,2,\ldots,12$. The main goal of the following scheme is to put in evidence the role played by changes of population size and the difference between the monthly index which does not consider these changes and the correct one. In the next formula (1) we use the population *y*-average monthly estimate given by: $\sum_{i \in s_t} y_{itm} / n_{tm} / = \tilde{\tilde{y}}_{tm}$, which holds for and year T and month m.

Table 1: *Index calculation for taking into account business demography*

General sampling design	**Simple random sampling (*srs*)**
$I_{Tm/T} = \dfrac{12\sum_{i\in s_T} y_{iTm} w_{iT}}{\sum_{m=1}^{12}\sum_{i\in s_T} y_{iTm} w_{iT}}$	$I_{Tm/T} = \dfrac{12\sum_{i\in s_T} y_{iTm}}{\sum_{m=1}^{12}\sum_{i\in s_T} y_{iTm}}$
$I_{T+1,m/T} = I_{Tm/T}\dfrac{\sum_{i\in s_{T+1}} y_{i,T+1,m} w_{i,T+1,m}}{\sum_{i\in s_{T+1}} y_{iTm}\hat{w}_{iTm}}$	$I_{T+1,m/T} = I_{Tm/T}\dfrac{\sum_{i\in s_{T+1}} y_{i,T+1,m}}{\sum_{i\in s_{T+1}} y_{iTm}}\left(\dfrac{N_{T+1}}{N_T}\right)$
$I_{T+k,m/T} = I_{T+k-1,m/T}\dfrac{\sum_{i\in s_{T+k}} y_{i,T+k,m} w_{i,T+k,m}}{\sum_{i\in s_{T+k}} y_{i,T+k-1,m}\hat{w}_{i,T+k-1,m}}$	$I_{T+k,m/T} = I_{T+k-1m/T}\dfrac{\sum_{i\in s_{T+k}} y_{i,T+k,m}}{\sum_{i\in s_{T+k}} y_{i,T+k-1,m}}\left(\dfrac{N_{T+k}}{N_{T+k-1}}\right)$

It is straightforward to obtain the general recursive formula under (*srs*):

$$I_{T+k,m/T} = I_{Tm/T}\prod_{h=1}^{k}\left(\frac{\sum_{i\in s_{T+h}} y_{i,T+h,m}}{\sum_{i\in s_{T+h}} y_{i,T+h-1,m}}\right)\left(\frac{N_{T+k}}{N_T}\right) \approx I_{Tm/T}\left(\frac{\tilde{\bar{y}}_{T+k,m}}{\tilde{\bar{y}}_{T,m}}\right)\left(\frac{N_{T+k}}{N_T}\right) =$$

$$= I_{Tm/T}\tilde{Y}_{T+k,m/Tm} D_{T+k/T} \qquad (1)$$

where $\tilde{Y}_{T+k,m/Tm}$ is the sample estimate of the *y*-change between the month *m* of the years (*T*+*k*) and *T*, and $D_{T+k/T}$ is the demographic index between years (*T*+*k*) and *T*.

A more detailed framework is characterized by the case when updates of the monthly population size are available on a monthly basis. In this situation, the index which includes monthly updates of the sampling weights under (*srs*) can be obtained using similar argumentations as above, so that we have the index recursive formula:

$$I_{T+k,m/T} = I_{Tm/T}\prod_{h=1}^{k}\left(\frac{\sum_{i\in s_{T+h}} y_{i,T+h,m} N_{T+h,m}}{\sum_{i\in s_{T+h}} y_{i,T+h-1,m} N_{T+h-1,m}}\right) \approx I_{Tm/T}\tilde{Y}_{T+k,m/Tm} D_{T+k,m/Tm} \qquad (2)$$

where $D_{T+k,m/Tm} = N_{T+k,m}/N_{Tm}$. The overall demographic impact on the *y*-variable changes between years (*T*+*k*) and *T* is given by $D_{T+k/T}$ if monthly population changes are ignored, while it can be approximated by: $\sum_{m=1}^{12} D_{T+k,m/Tm}/12$ when monthly population size updates can be taken into account. Broadly speaking, the longest is the time distance between actual reference month *m* and the base year, the largest may be the demographic effect, e.g. the index change due to business demography. Unless the demographic effect, the index changes would be due to changes of the *y*-average only.

In current practice, monthly updates of sampling weights because of business demography may be not realistic. Normally they may be applied if an external administrative source is able to produce infra-annual estimates of deaths and births, which may be included into the sampling weights monthly update scheme after a proper reconciliation with definitions, observation domain and periodicity (they may not refer to the whole month or to ordinary calendar months) used in the survey. On the other hand, the calculation scheme in the previous table implies that no demographic effect is taken into account for 12 consecutive months – until December each year – while it is included in monthly estimates starting from January next year:

as a consequence, a time series gap may occur at each January, which may affect longitudinal dynamics and seasonal adjustment. A methodology for smoothing the monthly weights update (Beaumont and Rivest, 2007) consists in using monthly sample information on deaths. The basic rationale is founded on the following steps.
For the whole year T we suppose to know monthly business demography, e.g. all the addenda of the following identity: $N_{Tm} = N_{T,m-1} + b_{T,m} - d_{T,m}$ for m=2,3,…,12, where b_m is the number of units born between month (m-1) and m and d_m is the number of units dead in the same period. Furthermore, we suppose to estimate a statistical model which explains new units (born) on the basis of dead units, for instance:

$$\hat{b}_{T,m} = \hat{\alpha}_1 N_{T,m-1} + \hat{\alpha}_2 d_{T,m}. \tag{3}$$

As soon as information on deaths is available – let's suppose starting from January year (T+1) – we can first calculate the sample estimate of the overall number of units dead in the whole population during January year (T+1). If $d_{i,T+1,1}$ is a binary variable equal to 1 if the *i-th* sample unit is dead during January year (T+1) and to 0 otherwise, under (*srs*) the previous estimate is given by:

$$\hat{\hat{d}}_{T+1,1} = \frac{N_{T+1,1}}{n_{T+1,1}} \sum_{i \in s_{T+1}} d_{i,T+1,1} \tag{4}$$

and we can use the estimate (4) into the recursive formula (3), applied to January year (T+1) using the same parameter estimates obtained through (3):

$$\hat{b}_{T+1,1} = \hat{\alpha}_1 N_{T,12} + \hat{\alpha}_2 \hat{\hat{d}}_{T+1,1}. \tag{5}$$

Finally, we can use (5) in order to estimate:

$$\hat{N}_{T+1,1} = N_{T,12} + \hat{b}_{T+1,1} - \hat{\hat{d}}_{T+1,1} \tag{6}$$

which can be used for updating the monthly sampling weights $\hat{N}_{T+1,1} / n_{T+1,1}$.
The general estimation scheme from February onward (m=2,3,…,12), is:

$$\hat{\hat{d}}_{T+1,m} = \frac{N_{T+1,m}}{n_{T+1,m}} \sum_{i \in s_{T+1}} d_{i,T+1,m} \tag{4'}$$

$$\hat{b}_{T+1,m} = \hat{\alpha}_1 N_{T+1,m-1} + \hat{\alpha}_2 \hat{\hat{d}}_{T+1,m-1} \tag{5'}$$

$$\hat{N}_{T+1,m} = N_{T+1,m-1} + \hat{b}_{T+1,m-1} - \hat{\hat{d}}_{T+1,m-1}. \tag{6'}$$

The previous monthly estimates can be used for updating monthly sampling weights. The main problem is that, of course, it may happen that, when the true population size referred to January next year is available, we have $\hat{N}_{T+2,1} \neq N_{T+2,1}$. If $\hat{N}_{T+2,1} = \beta N_{T+2,1}$, when β>1 (<1) there will be an overestimation (underestimation) of monthly weights.
A simple method for tackling the problem is as follows. We suppose to use all the monthly data on births and deaths during year (T+1) for estimating the previous gap parameter β – say $\hat{\beta}$. Starting from January year (T+2) we can use the new estimate :

$$\hat{N}_{T+2,m^*} = \hat{N}_{T+2,m} / \hat{\beta}. \qquad \text{for } m=1,2,\ldots 12 \tag{7}$$

The extension to the following years can be obtained straightforwardly.

3. Sampling new units from the updated business register

We suppose that information on new units (births: B) is available in January each year, even though in real contexts business demography it happens on a daily basis. In January year (T+1) the reference population size can be written as follows:

$$N_{T+1} = N_{T+1(O)} + N_{T+1(B)} \tag{8}$$

where $N_{T+1(O)}$ is the number of units which already existed at the beginning of year T (O: old units) – and which are supposed to be the only units known to be active during year T – while $N_{T+1(B)}$ is the number of units born during the year T, about which no statistical information is available from any source before January year (T+1). If the economic behaviour of the $N_{T+1(B)}$ new units is different from that of the pre-existing $N_{T+1(O)}$ units, it may be recommended to update the sample in year (T+1) in order to include a subset of new units and observe them. This re-sampling mechanism (Rao *et al*., 1992) consists in the application of a post-stratification level, in addition to the stratification scheme currently applied in the survey, where in each original stratum the additional split into 2 new strata will be based on the feature *Old* or *Born* to be assigned to each population unit. Each sample size related to the year (T+1) – for simplicity we suppose to use the same sample in all months in (T+1) – can be written:

$$n_{T+1} = n_{T+1(O)} + n_{T+1(B)} \tag{9}$$

and the calculation of indexes which implies the potential different pattern of new and old units is resumed in the following table 2, for year (T+1) and a generic year (T+k).

Table 2: *Index calculation for taking into account different profiles of new units*

Year	**Simple random sampling (*srs*)**
T+1	$I_{T+1,m/T} = I_{Tm/T}\left[\dfrac{\sum_{i\in s_{T+1(O)}} y_{i,T+1,m}\left(\dfrac{N_{T+1(O)}}{n_{T+1(O)}}\right) + \sum_{i\in s_{T+1(B)}} y_{i,T+1,m}\left(\dfrac{N_{T+1(B)}}{n_{T+1(B)}}\right)}{\sum_{i\in s_{T+1(O)}} y_{iTm}\left(\dfrac{N_T}{n_{T+1(O)}}\right)}\right]$
T+k	$I_{T+k,m/T} = I_{T+k-1,m/T}\left[\dfrac{\sum_{i\in s_{T+k(O)}} y_{i,T+k,m}\left(\dfrac{N_{T+k(O)}}{n_{T+k(O)}}\right) + \sum_{i\in s_{T+k(B)}} y_{i,T+k,m}\left(\dfrac{N_{T+k(B)}}{n_{T+k(B)}}\right)}{\sum_{i\in s_{T+k(O)}} y_{i,T+k-1,m}\left(\dfrac{N_{T+k-1}}{n_{T+k(O)}}\right)}\right]$

In order to implement the sampling design, the new units (B) sample size in the year (T+1) can be determined through the proportional allocation given by: $n_{T+1(B)} = N_{T+1(B)} n_{T+1} / N_{T+1}$. A limit of the previous table approach is that by definition we do not know any values of born units related to the previous year T, since all new units have been supposed to start activity in January year (T+1). As a consequence, there is advantage if one supposes to ask, in the monthly questionnaires, for past y-amount in month (m-1) instead of (m-12). In this case, starting from February each year and for a generic year (T+k), the second formula in table 2 may be written including in estimations at denominator new units data as well (k=2,3,...,12):

$$I_{T+k,m/T^*} = I_{T+k,m-1/T}\left[\frac{\sum_{i\in s_{T+k(O)}} y_{i,T+k,m}\left(\dfrac{N_{T+k(O)}}{n_{T+k(O)}}\right) + \sum_{i\in s_{T+k(B)}} y_{i,T+k,m}\left(\dfrac{N_{T+k(B)}}{n_{T+k(B)}}\right)}{\sum_{i\in s_{T+k(O)}} y_{i,T+k,m-1}\left(\dfrac{N_{T+k(O)}}{n_{T+k(O)}}\right) + \sum_{i\in s_{T+k(B)}} y_{i,T+k,m-1}\left(\dfrac{N_{T+k(B)}}{n_{T+k(B)}}\right)}\right] = ...$$

$$....= I_{T+k,m-1/T}\left[\tilde{Y}_{T+k,m/T+k,m-1(O)}\frac{\hat{Y}_{T+k,m-1(O)}}{\hat{Y}_{T+k,m-1}} + \tilde{Y}_{T+k,m/T+k,m-1(B)}\frac{\hat{Y}_{T+k,m-1(B)}}{\hat{Y}_{T+k,m-1}}\right] \tag{10}$$

where $\tilde{Y}_{T+k,m/T+k,m-1(\gamma)}$ and $\hat{Y}_{T+k,m-1(\gamma)}$ are, respectively, the sample estimates of: a) the y-change between the months m and (m-1) of year T+k; b) the y-total in month (m-1), year (T+k), for units whose kind is γ (γ=O,B), with $\hat{Y}_{T+k,m-1(O)} + \hat{Y}_{T+k,m-1(B)} = \hat{Y}_{T+k,m-1}$.

4. Empirical test and perspective conclusions

Since 2001, ISTAT (Italian National Statistical Institute) elaborates and releases quarterly indexes concerning turnover of the "Wholesale trade and commission trade sector" (classification NACE Rev.2 division 46). This economic sector is characterized by a huge number of enterprises, even though a process of concentration is ongoing. The sampling survey is based on stratified random sampling (7.500 units), where strata cross each other 9 economic activities and 3 employment classes. On the basis of elementary strata indexes, calculations of higher order indexes – among which the total wholesale trade one – are based on weighted means of lower order indexes, where weights derive from structural business statistics. The quarterly questionnaire asks for turnover of the reference quarter *q* and of the same quarter of the previous year (*q*-4). Each quarterly index is calculated through the scheme in the second column of table 1 – where quarters substitute months – but without taking into account demography, so that quarterly sample weights refer to population in the base year. The actual base year is 2010 and population size updates will be introduced in 2015.

The simulation study reported in table 3 covers the period 2005-2010 and is based on a sub-sample derived from the real quarterly samples, where outliers and not panel units have been excluded, in order to deal with more steady estimates of yearly changes, resumed in the column (1). In this framework the base year is 2005. Taking into account the yearly changes of the average turnover by enterprise, from 2005 to 2010 the index of change would be equal to 22,1%. If we consider information on population size (2), born units (3) and dead units (4), we can calculate the yearly demographic indexes (5). The indexes which consider the dynamics of both the average turnover by enterprise and the business demography are given in (6)=(1)x(4): along the five years 2005-2010 the real change of *y*-amount in the whole population would be quite lower, 12,7%, because in the same time lag the demographic effect was lower than one (0,923), since deaths were significantly higher than births. The difference between the indexes without and with demographic effect is -0,094, which is -7,675 in percent and about 23.177 million euro in absolute value (huge amount).

Studies in progress are investigating monthly indexes dynamics and other real populations, as well as the possibility to manage demography on a monthly basis.

Table 3: *Simulation study for the Italian wholesale trade turnover 2005-2010*

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
T	$\tilde{Y}_{T+k/T}$	*N*	Born	Dead	$D_{T+k/T}$	(1)x(4)	(6)-(1)	[(6)/(7)-1]100
2010	1,221	741.783	80.111	94.321	0,923	1,127	-0,094	-7,675
2009	1,254	755.993	82.184	103.925	0,941	1,180	-0,074	-5,907
2008	1,267	777.734	85.562	97.481	0,968	1,226	-0,041	-3,201
2007	1,203	789.653	88.526	95.195	0,983	1,182	-0,021	-1,717
2006	1,117	796.322	92.045	99.173	0,991	1,107	-0,010	-0,887
2005	1,000	803.450			1,000	1,000		

Source: elaboration on ISTAT data.

References

Beaumont J.F. and Rivest L.P. (2007) "A Weight Smoothing Method for Dealing with Stratum Jumpers in Business Surveys", *SSC Annual Meeting 2007*, Proceedings of the Survey Methods Section.

Dufour J., Gagnon F., Morin Y., Renaud M. and Särndal, C.E. (2001) "A Better Understanding of Weight Transformation Through a Measure of Change", *Survey Methodology*, 27, 97-108.

Rao J.N.K., Wu C.F.J., and Yue K. (1992) "Some Recent Work on Resampling Methods for Complex Surveys", *Survey Methodology*, 18, 209-217.