SMALL AREA ESTIMATION AND THE LABOUR MARKET IN LOMBARDY'S INDUSTRIAL DISTRICTS: A METHODOLOGICAL APPROACH[*]

by

Eleonora Bartoloni

Istat, Regional Directorate, Lombardy. Bartolon@istat.it

Abstract: The aim of this study is to provide estimates of employment and unemployment for local areas which are under-represented in the Labour Force Survey set up by the National Institute of Statistics. I analyse the relative performance of different traditional direct and indirect estimators borrowing strength from survey data. The local areas under investigation are Lombardy's industrial districts as recently re-defined by regional law. The adoption of direct estimators for the industrial districts should lead to large standard errors due to the small sample size of the areas under investigation. The adoption of indirect estimators, which are crucially based on larger sample sizes, is a necessary step in order to produce more efficient estimations at the local area level. The results, even though limited to a small group of small areas, point out the need for further research on small area estimation. In particular, as concerns the specific labour market stocks analysed in this paper, more attention must be paid to unemployment by introducing both model-based estimations, and by adopting more suitable auxiliary variables at the local area level.

Keywords: Small Area Estimation, Industrial Districts
JEL Classification: C15, C42, R23

[*] I gratefully acknowledge Alessandro Faramondi for helpful comments and suggestions on an earlier draft of this paper. I would like also to thank seminar participants at the workshop on "Conoscere il territorio: stime per piccole aree" Università Cattolica del Sacro Cuore, 13 Maggio 2005, Piacenza and at the "XXVI Conferenza di Scienze Regionali" Napoli, 17-19 ottobre 2005. Needless to say, the usual disclaimer applies.

1 INTRODUCTION

The demand for statistics at sub-national level has significantly increased in all industrialised countries in recent years. We observe, on the one hand, an increasing need for ever more detailed socio-economic information at the local area level, relevant for policy analysis and implementation; while on the other hand, the main sources for official statistics are national surveys which assume large geographical areas, usually delimited by administrative boundaries, as common domains of estimates. Estimations for small geographical areas based on national sample designs lead to unacceptable results because of the small sample size of the local domains which negatively affects the precision of estimates.

Nevertheless, the enlargement of the sample size at the national level, which could increase the availability of reliable statistics for small domains, is an expensive alternative, not feasible in a situation characterized by a growing lack of public resources. In this framework small area estimation techniques represent a possible alternative because they lead to improved precision in the estimates at the local area level without modifying the overall sample design.

The research carried out in ISTAT based on this methodological approach started at the end of the eighties. At the present time series estimates of labour market stocks and value added are produced at the Local Labour Market Area level (Faramondi A. and Piras M. G, 2002).

The estimation of labour market stocks at the local area level is the aim of this contribution. The data source is the Italian Labour Force Survey (LFS), which produces quarterly estimates of the labour market stocks at the national and regional levels and annual estimates at the provincial level. In particular, I carried out a Monte Carlo Simulation, applying the sample design of the LFS to Lombardy’s Population Census data in order to analyse relative performances in terms of bias and efficiency of different traditional direct and indirect estimators at Lombardy’s Industrial District level. Given the sample design of the LFS, the adoption of direct estimators for the industrial districts should lead to large standard errors due to the small sample size of the areas under investigation. The adoption of indirect estimators, which are crucially based on larger sample sizes, is a necessary step in order to produce more efficient estimations at the local area level.

2 THE TERRITORIAL UNITS UNDER INVESTIGATION

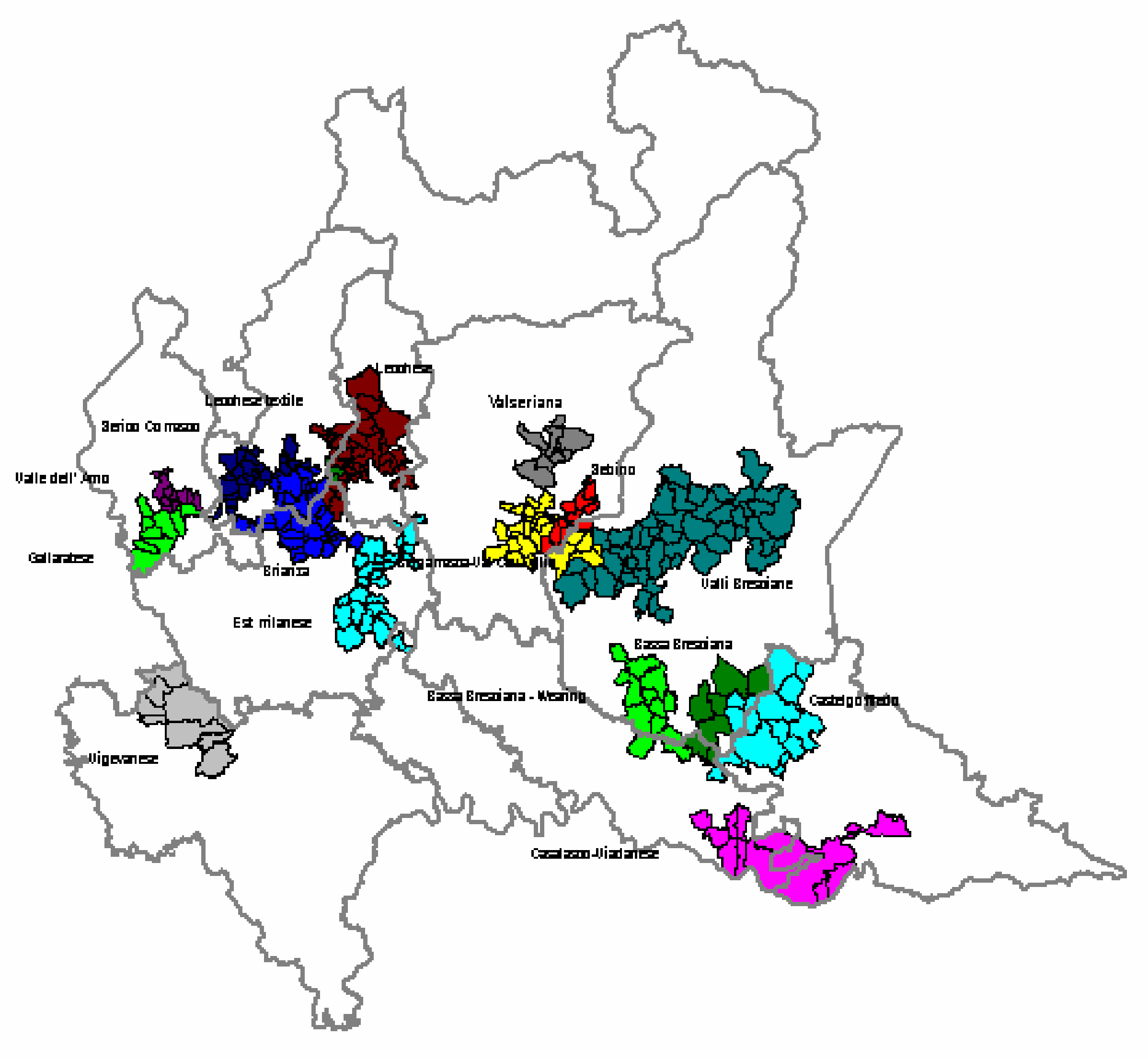
The local areas under investigation are Lombardy’s industrial districts, as recently re-defined by regional law. They may be defined as aggregations of neighbouring municipalities in which small and medium-sized firms specialising in manufacturing activities are backward and forward linked in the same production chain.

The spatial proximity of these firms represents a vehicle for interaction and knowledge transfer and, ultimately, an opportunity for the social and economic growth of the local area as a whole.

The 16 Industrial districts cover 302 municipalities localized in 10 provinces. Seven districts are specialized in textile-wearing activities, three in the metal products sector, two in the footwear sector, 1 in the furniture sector, 1 in wood products, 1 in electrical and professional goods and 1 in the rubber and plastic sector.

The localization of the industrial districts in Lombardy is shown in chart 1, while in table 1 some socio-economic indicators are given which pinpoint the main structural characteristics of such industrial clusters.

Chart 1 - Lombardy's industrial districts



Lombardy's industrial districts cover almost 20% of the regional area. Some districts, such as the Valli Bresciane and Casalasco Viadanese are large (more than 800 km^2 and 400 km^2 respectively), while others cover a small territorial area (e.g.: Lecchese-textile and Valle dell'Arno, cover less than 50 km^2,).

The population of these local areas (about 2,200,000) represents about 24% of the regional total, while population density shows different patterns across the districts, ranging from 1,000,800 inhab/km^2 in the district of Brianza to less than 140 inhab/km^2 in the Casalasco Viadanese district.

According to 2001 Census data, the local units in the industrial districts employed more than 870,000 people, corresponding to 26% of the regional total. About one half of the workers in the industrial districts are employed in manufacturing firms, which, in turn, represent about 35% of the manufacturing sector in Lombardy.

Tab.1 - Lombardy's Industrial Districts: structural characteristics

	Sector of	Area	Population	employees in the local units - 2001	
	specialisation		density	Total	of which: manufacturing
	(Ateco 1991)	km^2	inhab/km2		(%)
01 Valle dell'Arno	Metal products (27-28)	46,4	902	16.478	71,3
02 Lecchese	Metal products (27-28)	310,1	685	80.955	53,1
03 Valli Bresciane	Metal products (27-28)	801,0	387	116.947	56,5
04 Serico comasco	Textile (17)	171,6	1.123	77.400	38,1
05 Valseriana	Textile (17)	118,1	407	19.196	60,9
06 Castelgoffredo	Textile (17)	354,0	180	28.082	60,9
07 Bassa Bresciana	Leather and Footwear (19)	188,1	196	11.027	56,6
08 Sebino	Rubber and plastic (25)	81,3	505	18.542	62,5
09 Est Milanese	Electrical and Professional goods (31-32-33)	249,5	1.155	153.886	39,0
10 Brianza	Furniture (36)	258,4	1.808	152.280	47,8
11 Bergamasca-Val Cavallina-Oglio	Wearing and Furniture (18-36)	229,2	644	67.006	56,4
12 Lecchese (textile)	Textile (17)	32,6	811	12.114	68,3
13 Bassa Bresciana (wearing)	Wearing (18)	210,6	214	16.870	61,1
14 Gallaratese	Wearing (18)	124,7	1.066	50.934	44,4
15 Vigevanese	Leather and Footwear, Machinery and Equipment (19-29)	270,0	348	30.011	43,1
16 Casalasco Viadanese	Wood products exp. Furniture (20)	406,9	137	19.415	52,2
TOTAL DISTRICTS		**3.852**	**572**	**871.143**	**49,6**
OTHER AREAS		**16.214,8**	**434**	**2.511.269**	**31,4**
LOMBARDY		**20.067,1**	**461**	**3.382.412**	**36,1**

Source: Statistical Yearbook of Lombardy Region

3 SOME METHODOLOGICAL REMARKS

The estimation of employment and unemployment at the industrial district level is based mainly on sample units provided by the Labour Force Survey (LFS).

This survey provides quarterly estimates for labour market stocks at the national and regional level and annual estimates at the provincial level (ISTAT, 1991; Falorsi P.D. and Falorsi S., 1994).

In order to simplify the methodological explanation I introduce the following notation:

d $(d=1,\ldots.,D)$ is the generic industrial district;

$prov(prov=1,\ldots L_d)$ is the generic province including the district d

h $(h=1,\ldots,H_{d\,prov})$ is the stratum index referring to the LFS provincial-level stratification;

i $(i=1,\ldots,P_h)$ is the primary sampling unit (PSU) index;

j $(j=1,\ldots,S_{hi})$ is the secondary sampling unit (SSU) index;

a $(a=1,\ldots,A)$ is the index relative to the combination of sex and age class variables;

N_h is the total number of persons in the stratum h;

N_{hi} is the total number of persons in the PSU i of stratum h;

N_{hij} is the total number of persons in the SSU j of the PSU i of stratum h;

N_{hija} is the total number of persons in hij of sex-age group a.

The stratification is undertaken at the provincial level, the size in terms of population being the stratification criterion adopted. In each province the PSUs are divided into two groups according to their size. To the first group belong the so called “self-representing” municipalities, i.e. the larger PSUs . Each PUS is selected.

To the second group belong the smaller PSUs which are also called “non self-representing” municipalities. The selection of these PSUs is made within the stratum of similar size. From each stratum two PSUs are selected without replacement and with probability proportional to size in terms of population. Finally, the SSUs are selected without replacement and independently by each PSU extracted. Each SSU has an equal probability of being selected.

Hence, the parameter under investigation is:

$$Y_d = \sum_{a=1}^{A} \sum_{prov=1}^{L_d} \sum_{h=1}^{H_{dprov}} \sum_{i=1}^{P_h} \sum_{j=1}^{S_{hi}} Y_{ahij} \qquad 1)$$

A traditional direct estimator of Y_d is the well known *Horvitz-Thompson* estimator (HT) which assigns to each sample unit a basic weight expressed by[1]:

$$K_{hij} = \frac{N_h}{t_h N_{hi}} \frac{S_{hi}}{s_{hi}}$$

where s_{hi} is the number of sample households in the PSU *i* of stratum *h* and t_h is the number of PSUs extracted by each stratum.

According to the probability sampling scheme, the HT estimator is unbiased (Cochran, 1977, p.260), but it is not appropriate when the estimates refer to small domains which are not represented in the sampling design. In fact, the sampling design of the LFS is based on a provincial-based stratification of the sample units, and thus it does not support direct estimations for local area aggregations of the sample which cut across the planned domains.

In this context the domain sample size may be considered a random variable whose realized values may significantly vary from one sample to another, and it could also be zero.

A common way to reduce the natural variability of the HT estimator and to improve the accuracy of direct estimations is provided by the ratio estimator. In particular, the *post-stratified ratio* estimator (POS) is actually employed by ISTAT to derive labour market estimations at the provincial level, and is given by:

$${}_R\hat{Y}_d = \sum_{a=1}^{A} \frac{\hat{Y}_{da}}{\hat{N}_{da}} N_{da} \qquad 2)$$

where $\hat{Y}_{da}$ and $\hat{N}_{da}$ are the direct estimates based only on the sample units belonging to the sex-age group[2] *a* of the industrial district *d* and N_{da} is the total population. The POS estimator is defined as approximately unbiased (Rao, 2003, p.17), depending on the ratio $\frac{\hat{Y}_{da}}{\hat{N}_{da}}$ bias, which is negligible in large samples.

The peculiarity of this estimator, which is a simpler form of the *General Regression (GREG) estimator* (Rao, 2003, p.14), is that of borrowing straight from an auxiliary variable known at

[1] A modified version of the HT estimator is the *Calibration estimator*. Based on a calibration of the basic weights of the HT estimator, it represents a simple form of the *General Regression (GREG) Estimator* in the case of single auxiliary variable (Rao, 2003, p.20). Actually ISTAT employs the Calibration Estimator to derive quarterly estimates of employment and unemployment at the national and regional level.

[2] In this work the direct estimator employed for the POS is the HT estimator. I consider 4 post-strata, corresponding to the sex-age (40 years and under, over 40 years) combination. A deeper subset of the age-class variable would lead to set up post-strata, in many cases, without sample units.

the local area level and significantly correlated to the parameters under investigation. At present the only information with such characteristics, and available on a yearly basis, is the population by sex and age class drawn from the municipalities' registers.

In the POS expression the ratio $\frac{N_{da}}{\hat{N}_{da}}$ represents an adjustment factor which may contrast the variability of $\hat{Y}_{da}$ and for this reason the POS estimator should be more efficient than the HT estimator.

The adoption of direct estimates for the industrial districts leads to large standard errors due to the small sample sizes of the area under investigation where the PSUs (municipalities) and the SSUs (households) are very low in absolute values.

Table 2 reports the total population and the characteristics of the 2003 LFS units for Lombardy and its industrial districts. This shows that 6 industrial districts have only 1 PSU and that two industrial districts (Valseriana and Bassa Bresciana – leather and footwear), have a null sample size, rendering direct estimations unavailable.

Tab.2 - Total Population and Labour Force Survey units in the Lombardy Region – 2003

		Municipalities (PSUs) in the LFS			Population in the LFS	
	Total number of Municipalities	*number*	*% on PSUs in Lombardy*	Total Population	*number*	*% on LFS population in Lombardy*
01 Valle dell'Arno	11	2	1,0	41.465	776	0,9
02 Lecchese	40	8	4,2	210.277	3.651	4,1
03 Valli Bresciane	49	7	3,7	303.735	2.934	3,3
04 Serico comasco	26	4	2,1	110.404	1.632	1,9
05 Valseriana	10	-	-	47.682	-	-
06 Castelgoffredo	15	4	2,1	62.492	1.540	1,7
07 Bassa Bresciana	8	-	-	36.294	-	-
08 Sebino	11	1	0,5	40.255	476	0,5
09 Est Milanese	28	7	3,7	284.889	2.882	3,3
10 Brianza	36	10	5,2	459.611	4.132	4,7
11 Bergamasca-Val Cavallina-Oglio	26	4	2,1	144.314	1.925	2,2
12 Lecchese (textile)	9	1	0,5	25.997	469	0,5
13 Bassa Bresciana (wearing)	12	1	0,5	44.265	329	0,4
14 Gallaratese	9	1	0,5	130.759	179	0,2
15 Vigevanese	8	1	0,5	91.695	550	0,6
16 Casalasco Viadanese	13	1	0,5	54.785	369	0,4
TOTAL DISTRICTS	**311**	**52**	**27,2**	**2.088.919**	**21.844**	**24,8**
OTHER AREAS	**1.245**	**139**	**72,8**	**7.055.591**	**66.177**	**75,2**
LOMBARDY	**1.556**	**191**	**100,0**	**9.144.510**	**88.021**	**100,0**

As a whole, the PSUs localised in the industrial districts represent only 27% of the total number of PSUs in the region. In terms of individuals, the sample units in the industrial districts correspond to one quarter of the regional sample units. The larger districts in terms of sample size are the Brianza, Lecchese, Est-Milanese and Valli Bresciane districts: in terms of PSUs their percent share on the Lombardy sample ranges between 5.2% and 3.7%.

Taking this framework into account, the adoption of indirect estimators, which are crucially based on larger sample sizes, is a necessary step in order to produce more efficient estimations at the local area level.

In this contribution I firstly formulate different indirect estimators, then I assess their reliability comparing them to the direct estimators.

This analysis is based on a Monte Carlo Simulation conducted on the 1991 Census Data. The simulation consists in the application of the sample design of the LFS in Lombardy to Lombardy's Census data. The design is a two-stage sampling with stratification of the PSUs. The samples selected at every replication consist of 175 PSUs from a population of 1,546 Municipalities and of 8,948 SSUs from a population of almost 3,000,300 households. I have extracted 1000 samples independently [3].

3.1 Indirect estimates of employment and unemployment: the synthetic estimator

The general formulation of the Synthetic Estimator is given by:

$$_{S}\hat{\boldsymbol{Y}}_{\boldsymbol{d}} = \sum_{\boldsymbol{a}=1}^{\boldsymbol{A}} \frac{\hat{\boldsymbol{Y}}_{\boldsymbol{ra}}}{\boldsymbol{N}_{\boldsymbol{ra}}} \boldsymbol{N}_{\boldsymbol{rda}} \qquad 3)$$

where $\hat{\boldsymbol{Y}}_{\boldsymbol{ra}}$ is the direct estimator of the sex-age group[4] *a* in the macro-area *r* where district *d* is localised, $\boldsymbol{N}_{\boldsymbol{ra}}$ is the known population of the sex-age group *a* in the macro-area *r* and $\boldsymbol{N}_{\boldsymbol{rda}}$ is the known population in the sex-age group *a* of district *d*.

The synthetic estimator given by 3) is a special case of the *regression-synthetic estimator* (Rao, 2003. p.46) and assumes that the small area and the macro-area are characterised by the same mean values of the parameter in each sex-age group under investigation. This assumption is crucial; in fact, when the district characteristics in terms of mean values in each sex-age group differ from those of the macro-area, the estimations are strongly biased. By considering the ratios: $\boldsymbol{R}_{\boldsymbol{ra}} = \frac{\boldsymbol{Y}_{\boldsymbol{ra}}}{\boldsymbol{N}_{\boldsymbol{ra}}}$ and $\boldsymbol{R}_{\boldsymbol{da}} = \frac{\boldsymbol{Y}_{\boldsymbol{da}}}{\boldsymbol{N}_{\boldsymbol{da}}}$, the bias of the Synthetic estimator may be expressed as:

$$\sum_{\boldsymbol{a}=1}^{\boldsymbol{A}} (\boldsymbol{R}_{\boldsymbol{ra}} - \boldsymbol{R}_{\boldsymbol{da}}) \boldsymbol{N}_{\boldsymbol{rda}} \qquad 4)$$

In this framework, the choice of the reference macro-area is crucial to get efficient estimations minimising expression 4). In this paper I selected 4 different formulations of the Synthetic Estimator corresponding to different clustering hypotheses.

The first formulation takes the region as the reference area. The advantage of this estimator, which is denoted as $_{S_1}\hat{\boldsymbol{Y}}_{\boldsymbol{d}}$, is that it takes the entire set of sample observations into account,

[3] The samples extracted are not of identical size in terms of PSUs and SSUs because of the partial cut-offs of 3 pairs of districts. The random extraction of a PSU belonging to two overlapped areas has determined the duplication of the PSU in order to include the Municipality in both the overlapped districts.

[4] For the Synthetic Estimator I consider 28 sex-age groups, corresponding to 14 age classes.

leading to a lower variance of the estimations. On the other hand, if each district exhibits different patterns with respect to Lombardy as a whole in term of mean values of the parameters in each sex-age groups, estimations based on this formulation may be heavily biased, with bias given by 4). The second formulation (${}_{S2}\hat{\boldsymbol{Y}}_{\boldsymbol{d}}$) takes groups of provinces as the reference area and is based on the assumption that an industrial district and the provinces to which it belongs exhibit homogeneous patterns in each sex-age group.

Tab.3 - The distribution of Lombardy's Industrial Districts across the provinces

Number of municipalities

Industrial districts	Varese	Como	Sondrio	Milan	Bergamo	Brescia	Pavia	Cremona	Mantua	Lecco	Lodi
01 Valle dell'Arno	11										
02 Lecchese		7		4	1					28	
03 Valli Bresciane						49					
04 Serico comasco		26									
05 Valseriana					10						
06 Castelgoffredo						3		1	11		
07 Bassa Bresciana						6		2			
08 Sebino					10	1					
09 Est Milanese				24	2					1	1
10 Brianza		16		20							
11 Bergamasca-Val Cavallina-Oglio					22	4					
12 Lecchese (textile)		2								7	
13 Bassa Bresciana (wearing)						10		2			
14 Gallaratese	9										
15 Vigevanese							8				
16 Casalasco Viadanese								8	5		
Total districts	**20**	**51**		**48**	**45**	**73**	**8**	**13**	**16**	**36**	**1**

Because 10 districts out of 16 span over more than one province (see table 3), the formulation of this estimator requires an intermediate step in order to split the districts in "sub-districts", each of which is entirely localised in one province. Next, I have derived the estimations for each of the resulting 31 sub-districts, by assigning to each sex-age group the mean values of the direct estimations resulting from the appropriate province. The synthetic estimation at the district level is finally obtained by sub-district aggregation.

Estimations based on this synthetic estimator have a higher variability because of the smaller number of sample units compared to the estimations based on the regional macro-area. On the other hand, the bias could be lower if the labour market in Lombardy exhibited a provincial characterisation instead of a regional one.

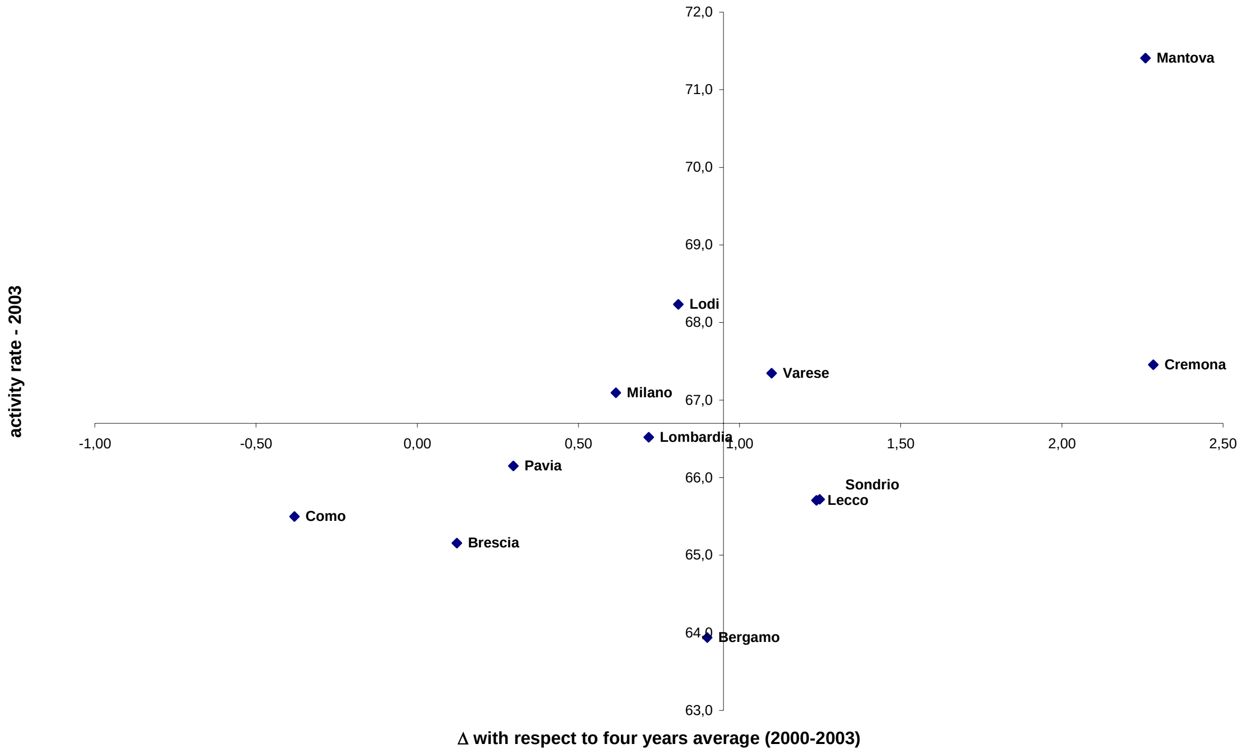
Similarly, for the third type of synthetic estimator (${}_{S3}\hat{\boldsymbol{Y}}_{\boldsymbol{d}}$) I have derived the related macro-areas as groups of provinces, but now the clustering proposed takes into consideration the relationship between the level of the provincial activity rate and its change during the period 2000-2003. As a measure of change in the labour market participation rate I consider the 2000-2003 *delta* determined as the 2003-2000 percent difference with respect to the four-year average. The source of the provincial activity rates is the LFS.

In figure 2 the intersection of the horizontal and vertical axes corresponds to the mean values, across Lombardy's provinces, of the 2000-2003 *delta* and activity rate. Hence, the points in the graph must be interpreted as shifts from mean values. Thus it is possible to classify at least three groups of provinces. The first group includes Milan and Lodi, with activity rates higher than the mean value, but with a performance during the period 2000-2003 which is lower, in terms of variation, than the average.

The second group includes Mantua, Cremona and Varese, where both activity rates and its variation is higher with respect to their relative mean values.

The third group includes the remaining six provinces whose variation (Δ) is more spread around the mean value, but whose activity rates stand below the mean value.

Fig. 2 - Activity rate and its variation by province



Also this formulation of the synthetic estimator (${}_{S3}\hat{Y}_d$) is characterised by a higher variability with respect to the synthetic estimator based on the regional macro-area. Nevertheless the efficiency of this estimator is crucially affected by the bias expressed by equation 4).

Finally, for the fourth expression of the synthetic estimator I introduce a different definition of the reference macro-area. The idea behind this definition is that of considering the economic performances of the productive systems localised in the industrial districts, on the basis of which clusters of municipalities can be derived. In this case, clusters are not based on administrative boundaries but on the interpretation of underlying economic patterns. This analysis is based on the examination of the intercensual change of employment in the sectors of specialisation by industrial districts.

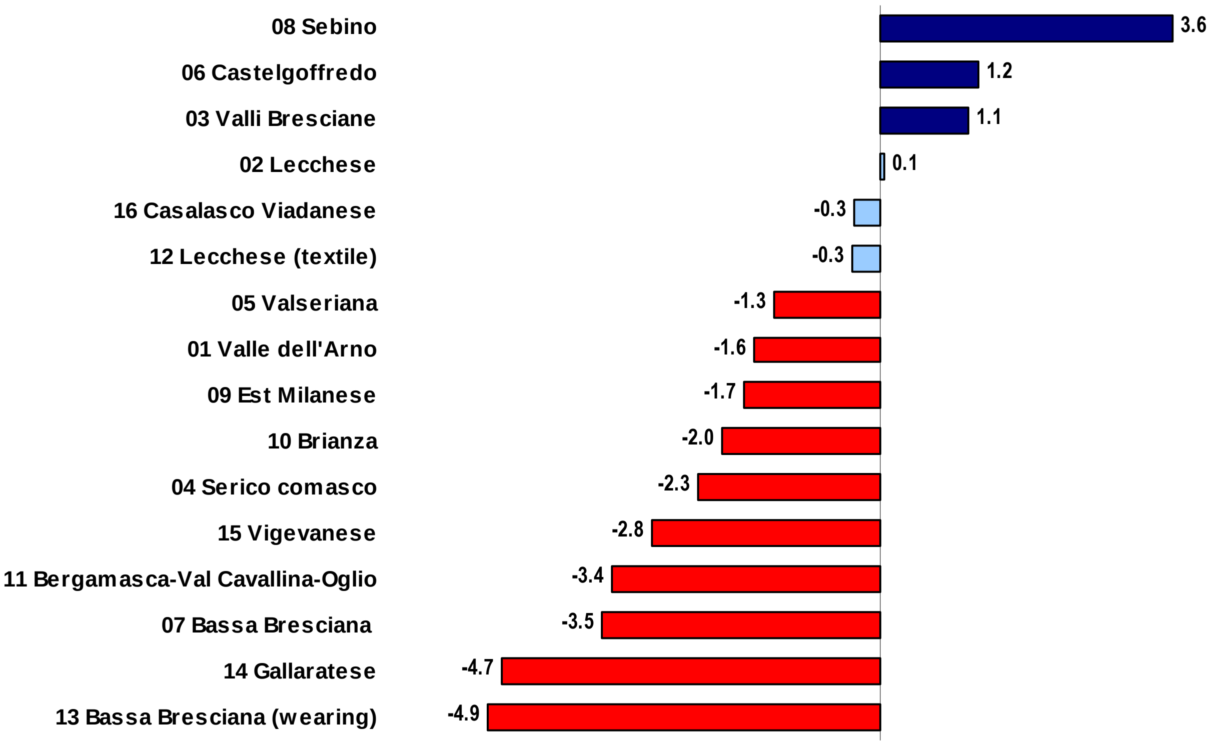
As shown in figure 3, only three districts have been characterised by a positive annual employment growth rate in the sectors of specialisation (Sebino, Castelgoffredo and the Valli Bresciane). Three other districts show rather limited variation (Lecchese, Casalasco Viadanese and Lecchese-Textile) and the remaining 10 districts have performed negatively during this period, as the annual growth rate of employment ranges from –1.3 % in the Valseriana district to –4.9% in the Bassa Bresciana-textiles district.

I have thus selected three clusters of municipalities (increasing employment, stationary employment and decreasing employment) corresponding to different employment increase/decrease rates in the manufacturing sector. These clusters represent the reference macro-area for the synthetic estimator ${}_{S4}\hat{\boldsymbol{Y}}_d$.

I would expect a higher variability of ${}_{S4}\hat{\boldsymbol{Y}}_d$ than ${}_{S3}\hat{\boldsymbol{Y}}_d$; in fact, for the last definition of macro-area the clusters of municipalities correspond to groups of PSUs and SSUs localised within the Industrial Districts, which in turn represent, as previously showed, only one quarter of the individuals in the LFS Lombardy sample. On the other hand this formulation could produce more efficient estimations if the ratio differences by sex-age groups expressed by the 4) were small.

Fig.3 - Intercensual change of employment in Lombardy's Industrial Districts

average annual % change in the local units' employment rate in the sectors of specialisation



3.2 Composite estimations: the Optimal Estimator and the Sample Size Dependant Estimator

Direct estimations of employment and unemployment based on the LFS sample design may be highly unstable, as stressed in the previous sections, because of the expected low number of PSUs and SSUs in the domains.

On the other hand, the synthetic estimator is more precise but may be biased if the small areas under investigation exhibit strong individual patterns with respect to the associated macro-area.

The composite estimator operates a balance between these two opposite characteristics in order to carry out more efficient estimations. It is obtained as a linear combination of a direct estimator ($\hat{Y}_d$) and a synthetic estimator (${}_S\hat{Y}_d$). Therefore the composite estimator is given by:

$$ {}_C\hat{Y}_d = \alpha_d \hat{Y}_d + (1-\alpha_d)\,{}_S\hat{Y}_d \qquad 5) $$

with weight $(0 \le \alpha_d \le 1)$.

The optimal weight for α_d is obtained by minimising the MSE of the composite estimator with respect to α_d, and is given by:

$$ \alpha^*_d \approx \frac{MSE({}_S\hat{Y}_d)}{MSE({}_S\hat{Y}_d) + MSE(\hat{Y}_d)} \qquad 6) $$

It follows from 6) that when the MSE of the post-stratified ratio estimation is low compared to the synthetic one, the optimal weight will be close to one, assigning more preference to the direct estimator. Conversely, the optimal weight is close to zero and the composite estimator reduces to the synthetic one. Besides, the composite estimator reduces to the synthetic one even when the direct estimation is missing for a given small area.

A different form of the composite estimator assigns a common weight to Lombardy's districts as a whole. In this case the optimal weight is given by (Rao, 2003, p. 58):

$$ \alpha^* \approx \frac{\sum_{d=1}^{D} MSE({}_S\hat{Y}_d)}{\sum_{d=1}^{D} MSE({}_S\hat{Y}_d) + \sum_{d=1}^{D} MSE(\hat{Y}_d)} \qquad 7) $$

It is worth noting that the common weight of 7) minimises the total MSE of Lombardy's districts, ensuring efficient estimations for the industrial districts as a whole but the efficiency is not sure for each district.

The Sample Size Dependant (SSD) estimator is a composite estimator as well, but the weight α_d is defined as:

$$ \alpha_d = \begin{cases} 1 & if \;\; \hat{N}_d \ge \lambda N_d \\ \hat{N}_d / (\lambda N_d) & otherwise \end{cases} \qquad 8) $$

The value of the weights α_d depends only on $\hat{N}_d$ and N_d though the parameter λ controls the contribution of the synthetic estimator in 5). The idea behind this is that of assigning more weight to the direct estimator when the proportion of the realised sample size in the small area is large. On the contrary, when the sample size of the domain is small, the synthetic estimator is more reliable and its contribution increases.

In this work I consider different values of the parameter λ in order to obtain the best SSD estimation relative to the smaller MSE.

4 ESTIMATION PERFORMANCE

The performance of the estimators discussed in the previous section will be evaluated with respect to employment and unemployment counts, considering both bias and efficiency. The bias of the estimations will be evaluated in terms of Absolute Relative Bias (ARB), given by:

$$\boldsymbol{ARB} = \frac{1}{\boldsymbol{D}}\sum_{d=1}^{D}\left|{}_{d}\boldsymbol{ARB}\right| \qquad 9)$$

where the percentage relative bias in each district is given by:

$$\left|{}_{d}\boldsymbol{ARB}\right| = \left|\frac{1}{\boldsymbol{R}}\sum_{r=1}^{R}\left(\frac{{}_{d}\hat{\boldsymbol{Y}}_{r}}{{}_{d}\boldsymbol{Y}}-1\right)\right|100 \qquad 9.a)$$

where R is the number of simulations (R=1000) performed, D is the number of Lombardy's districts (D=16), ${}_{d}\hat{\boldsymbol{Y}}_{r}$ is the value of the estimator in the district d for the simulation *r* and ${}_{d}\boldsymbol{Y}$ is the "true" value of the parameter obtained from the Population Census[5].

The efficiency will be evaluated in terms of Relative Root Mean Squared Error (RRMSE) and is given by:

$$\boldsymbol{RRMSE} = \frac{1}{\boldsymbol{D}}\sum_{d=1}^{D}{}_{d}\,\boldsymbol{RRMSE} \qquad 10)$$

where the percent relative MSE for each Lombardy district is given by:

$${}_{d}\,\boldsymbol{RRMSE} = \frac{\sqrt{{}_{d}\,\boldsymbol{MSE}}}{{}_{d}\boldsymbol{Y}}100$$

and

$${}_{d}\,\boldsymbol{MSE} = \frac{1}{\boldsymbol{R}}\sum_{r=1}^{R}\left({}_{d}\hat{\boldsymbol{Y}}_{r}-{}_{d}\boldsymbol{Y}\right)^{2}$$

[5] It is worth noting that the LFS and the Population Census do not adopt the same definitions of employed and unemployed, so that the stocks resulting from these statistical sources do not coincide. Nevertheless, the aim of this simulation study is that of evaluating the relative performance of different small area estimators applying the LFS sample design to the Population Census and assuming that the employment and unemployment counts from the Census source are the "true" value of the parameter under investigation.

4.1 Performance for the Industrial Districts as a whole

In Table 4 I show the values obtained by expressions 9) and 10) with respect to the direct and indirect estimators previously described. It is worth noting that the reliability of direct estimations is crucially affected by the size of the parameter under investigation. In practise, as regards the labour market stocks, I expect that direct estimates of unemployed counts will show larger sample errors than the estimates of the employed. This fact may, in turn, affect the quality of the estimations for Lombardy's industrial districts where the small territorial extension (in terms of small number of primary sampling units) is combined with unemployment rates traditionally lower respect to the regional average. In some cases, the bad performance of the direct estimates of unemployed counts inevitably affects the reliability of the composite estimators.

Tab.4 - Percent Average Absolute Bias (%ARB) and Percent Average Root Mean Square Error (%RRMSE) - Employment and Unemployment

ESTIMATOR	employment		unemployment	
	ARB%	RRMSE%	ARB%	RRMSE%
HT	6.2	63.1	58.3	70.4
POS	25.2	37.8	66.2	70.8
SYN1	4.4	4.5	37.3	37.5
SYN2	4.1	5.2	48.0	49.1
SYN3	4.3	4.7	47.6	48.0
SYN4	4.8	17.8	48.9	51.0
COM1	4.2	4.5	38.3	38.8
COM2	4.5	4.8	41.8	43.1
SSD1 λ =¾	6.5	13.7	52.0	57.4
SSD2 λ =1	5.9	12.0	50.7	55.7
SSD3 λ =1,5	5.4	9.9	48.3	52.2
SSD4 λ =2	5.1	8.4	46.2	49.0

In this section I analyse, the overall performance of the complete set of estimators in terms of average over the 16 industrial districts of $\left|{}_{d}\boldsymbol{ARB}\right|$ and ${}_{d}\boldsymbol{RRMSE}$ firstly with respect to the employed and then with respect to the unemployed counts.
With respect to the employment counts the following findings are noteworthy:

- The average bias of the HT estimator is about one fourth lower than the bias of the POS estimator. It is worth noting that it is possible to reduce this bias to a half by excluding one district (coded by 12) where the average bias on the total number of replications is particularly high. This is due to the low sampling fraction of this district, which, in turn, determines a very small domain sample sizes at each replication.

- The bias of the POS estimator crucially depends on its definition: in fact in expression 2) the bias of $\hat{N}_{da}$ adds up to that of $\hat{Y}_{da}$, raising the bias of POS with respect to HT. Nevertheless POS estimator is more precise: its RRMSE is about one half of that of the HT estimator (see also: Cicchitelli, Herzel, Montanari, 1992, pp.112-119). The higher efficiency of POS confirms that the adoption of auxiliary information in the form of population by sex-age groups does reduce the variability with respect to the HT estimator. For this reason I use the POS estimator as direct component of the composite estimator.

- The synthetic estimator performs better than POS in terms of relative bias, by considering Lombardy’s industrial districts as a whole. In fact, the formulations used, according to four reference macro-area proposed, present average percent biases which range from 4.1% for the estimator with provincial macro-area (SYS2), to 4.8% for that with macro-areas corresponding to groups of industrial districts (SYS4).
- The good performance of the synthetic estimator in terms of percent bias means that the employment variable in Lombardy’s districts considered as a whole does not exhibit strong individual effects with respect to the regional pattern. This is true both when the reference macro-areas are defined on administrative criteria, and when the boundaries are defined considering specific socio-economic patterns at a sub-regional level.
- The analysis of the overall efficiency suggests that the synthetic estimator with regional macro-area (SYS1) performs better than the other synthetic estimators proposed. This evidence means, on the one hand, that the large expected sample size at the regional level leads to a reduction of the variability of the estimates, and on the other hand that the regional characteristics, in terms of mean values of employment in each sex-age group, don’t significantly differ from those of the industrial districts as a whole. This result seems to be in accordance with the low efficiency of the direct estimators which present, on the whole, high variability due to the small expected sample size of the industrial districts. For the POS estimator this variability adds up to the high bias previously observed.
- Among the synthetic formulations, SYN4 shows the poorer performance in terms of percent RRMSE (17.8%). Seemingly, this is due to the low expected number of sample units in the related macro-area which amplifies the variability with respect to the other synthetic estimators.
- Bearing in mind the bad results in term of average bias of the POS estimator, it seems reasonable to expect an insignificant improvement in the results with the composite estimators, at least for the industrial districts as a whole. Because of its higher efficiency among the synthetic estimators, I use SYN1 in the construction of the composite estimators. Between the two composite estimators proposed (COM1 and COM2), COM1 shows, in general, the lower bias and the higher efficiency. Nevertheless, this estimator doesn’t lead to significant improvement with respect to SYS1 in terms of average efficiency.
- Among the composite estimators, SSD shows the poorer performance both in terms of bias and efficiency. This result is somewhat as expected, as the sample rates of Lombardy’s industrial districts are quite low and this is a condition which is not favourable to the SSD estimator. In this analysis I consider different forms of the SSD estimator, each of which corresponds to different choices of λ, the parameter which checks for the contribution of the synthetic component. The best performance both in terms of bias and efficiency is obtained with λ=2. Nevertheless, even by considering this option, the ARB and RRMSE percentages are above the values of the SYN1 estimator.

Turning to the unemployed estimation, the analysis undertaken on the entire group of districts suggests the following considerations:

- Highly biased direct estimation results. In particular, as for the employment counts, POS is more biased than HT, depending on the high bias which characterizes the ratio in expression 2). In addition, POS doesn’t show improvements in terms of efficiency with respect to the HT estimator.
- Among the four synthetic estimators proposed, SYS1 is, as expected, the most efficient, then it is used as indirect component in the composite estimators. On the other hand, POS is used as the direct component, even if it does not show improvements in terms of precision with respect to HT.
- Among the composite estimators proposed, COM1 presents, on the whole, the less negative performance both in terms of bias and overall efficiency, but is less precise than SYS1. This is due to the high RRMSE of the POS component where, in turn, the high ARB represents the most relevant component.
- Among the SSD estimators, the SSD with λ=2 gives the better performances both in terms of bias and efficiency. Nevertheless, even for the unemployed counts, this estimator fits poorly, compared to the SYS and the COM1 estimators, both in terms of bias and efficiency.

4.2 Performance by Industrial District

In this section I consider for each industrial district the performance of a restricted group of estimators. I calculate three different percent ratios in order to provide more elements to evaluate the relative performances of the direct estimations, which are crucially affected by the sampling rates. The first ratio (*rapp_1* in table 5a) is the share of each industrial district’s population to the population of the strata including the industrial district under consideration. Bearing in mind the sample design, one would expect that the higher this ratio the lower would be the bias of the direct estimations. The second ratio (*rapp_2* in table 5b) indicates the district’s weight on the regional population. This ratio shows the relative size of the district: one would expect that the higher this ratio, the higher would be the expected sample size and , the precision of the direct estimates. The third ratio (*rapp_3* in table 5b) is the share of the population of the strata, including the specific district, to the regional population. One would expect that the higher the ratio the higher would be the precision of the estimates (if rapp_2 is not low).

Considering at first the bias of the employed counts, Table 5a shows that:

- The HT estimator presents a lower variability in terms of percent bias with respect to POS: ranging from 0.1% to 53%, its average value drops to 1.3% if we exclude 3 districts (7, 11 and 12) which shows bias values higher than those of the other districts. It is worth noting that the highest percent bias (53%) characterizes the district with the lowest value of rapp_1 (0.6%).
- The POS estimator shows a higher degree of variability in terms of relative bias, ranging from 0.1% to 81.8%. In addition, the relative bias is negatively correlated to rapp_1, and thus lower percent bias values (under 2%) characterizes those districts showing the higher sampling rates (see districts 3, 9, 10, 14 and 15). Higher bias values (more than 50%) are associated instead to lower rapp_1 values (see districts 7, 8 and 12).
- As concerns the precision of the direct estimations, percent RRMSE seems to be negatively correlated to rapp_2 and, even if to a lesser extent, to rapp_3. In particular, with reference to the POS estimator precision is higher (RRMSE less than 5%) in those industrial districts where rapp_2 is not less than 3%. Districts with rapp_2 index less than 1% present, instead, the higher percent RRMSE (values higher than 30%).

Tab.5a – Employment: Percent Absolute Relative Bias (%ARB) by industrial district and estimator

	rapp_1	HT	POS	SYN1	SYN2	SYN3	SYN4	COM1
01 Valle dell'Arno	2.5	1.4	44.6	6.7	4.0	3.7	6.6	6.6
02 Lecchese	3.2	1.4	13.5	1.6	4.3	2.9	4.2	1.7
03 Valli Bresciane	5.4	0.7	0.1	0.7	0.7	0.7	0.9	0.6
04 Serico comasco	2.7	2.3	6.0	6.2	3.5	7.5	6.1	5.8
05 Valseriana	1.4	0.6	34.0	5.4	7.6	6.8	5.4	5.4
06 Castelgoffredo	1.4	4.6	18.4	9.4	4.7	6.9	9.4	9.6
07 Bassa Bresciana (cuoio calzature)	1.3	10.3	59.6	4.0	4.9	4.6	3.9	4.0
08 Sebino	1.3	0.1	52.5	4.4	5.8	5.3	3.8	4.4
09 Est Milanese	11.8	0.5	1.7	1.6	1.4	1.8	1.5	1.6
10 Brianza	8.6	1.9	0.3	2.0	1.2	2.6	1.9	1.7
11 Bergamasca-Val Cavallina-Oglio	3.2	18.4	27.5	3.1	4.7	4.2	2.9	3.2
12 Lecchese tessile	0.6	53.0	81.8	4.0	6.1	5.4	6.6	4.1
13 Bassa Bresciana (confezioni abbigliamento)	1.8	0.6	40.0	5.5	6.9	6.7	5.6	5.6
14 Gallaratese	3.6	0.8	0.5	5.8	3.2	2.8	5.8	3.7
15 Vigevanese	2.7	1.6	0.6	0.1	1.5	0.0	0.4	0.1
16 Casalasco Viadanese	1.5	0.6	21.8	9.1	5.7	6.2	11.8	9.6
Total Industrial districts		**6.2**	**25.2**	**4.4**	**4.1**	**4.3**	**4.8**	**4.2**
Pearson correlation coeff. with rapp_1		*-0.28*	*-0.61*					

Tab.5b – Employment: Percent Relative Root Mean Squared Error (%RRMSE) by industrial district and estimator

	rapp_2	rapp_3	HT	POS	SYN1	SYN2	SYN3	SYN4	COM1
01 Valle dell'Arno	0.4	17.3	106.5	67.4	6.7	5.2	4.3	13.7	6.7
02 Lecchese	2.3	69.6	29.1	15.8	1.8	5.5	3.2	24.5	2.0
03 Valli Bresciane	3.1	56.8	29.8	4.6	1.2	3.0	1.8	27.0	1.2
04 Serico comasco	1.1	41.2	55.8	26.0	6.2	5.3	7.7	13.6	5.9
05 Valseriana	0.5	37.3	93.6	59.4	5.5	8.1	7.0	13.3	5.5
06 Castelgoffredo	0.6	43.9	57.3	32.9	9.5	5.7	7.2	26.0	9.7
07 Bassa Bresciana (cuoio calzature)	0.4	28.7	106.4	74.5	4.1	5.5	4.9	13.0	4.2
08 Sebino	0.4	30.3	110.0	70.1	4.5	6.4	5.5	26.0	4.5
09 Est Milanese	3.0	25.6	29.5	4.8	1.8	2.0	2.2	12.8	1.9
10 Brianza	4.9	57.3	21.7	4.9	2.2	1.9	2.8	12.8	2.1
11 Bergamasca-Val Cavallina-Oglio	1.2	37.6	52.3	43.4	3.3	5.2	4.5	12.9	3.3
12 Lecchese tessile	0.1	25.7	96.4	87.9	4.1	7.0	5.6	24.5	4.2
13 Bassa Bresciana (confezioni abbigliamento)	0.5	26.5	90.9	61.9	5.6	7.4	6.9	13.4	5.6
14 Gallaratese	1.4	38.6	44.6	7.2	5.9	4.6	3.6	13.5	4.7
15 Vigevanese	1.0	38.7	33.0	7.7	1.0	4.7	1.4	12.8	1.0
16 Casalasco Viadanese	0.6	38.4	52.8	37.0	9.2	6.2	6.6	25.1	9.7
Total Industrial districts			**63.1**	**37.8**	**4.5**	**5.2**	**4.7**	**17.8**	**4.5**
Pearson corr. coeff. with rapp_2			-0.76	-0.75					
Pearson corr. coeff. with rapp_3			*-0.65*	*-0.60*					

- As concerns synthetic estimations, SYS1, SYS2 and SYS3 result more efficient than direct estimations in every industrial districts. SYS4 results more efficient than POS in 10 districts to 16, while it shows a better performance with respect to HT in every district.

- SYS1 shows the higher performance with respect to the other synthetic estimators in 10 districts out of 16 but in specific cases it is possible to gain higher efficiency, without increasing bias, by introducing alternative formulations of the synthetic estimator (e.g., SYN2 for districts 4, 6 and 10, SYN3 for districts 1 and 14), thus signalling that specific composite estimators for each industrial district may be studied by applying different definitions of the reference macro-areas.
- The adoption of the COM1 estimator does not increase the efficiency of the estimations, being the direct component of this estimator generally more biased than SYS1. Only in four cases COM1 results less biased than SYS1 (districts 3, 4, 10 and 14) and in three cases (4, 10 and 14) COM1 results the most efficient estimator, while in the remaining district COM1 and SYS1 present the same percent RRMSE.

Tab.6 – Unemployment: Percent Relative Root Mean Squared Error (%RRMSE) and Percent Absolute Relative Bias (%ARB) by industrial district and estimator

	%ARB					%RRMSE					
	rapp_1	**HT**	**POS**	**SYN1**	**COM1**	*rapp_2*	*rapp_3*	**HT**	**POS**	**SYN1**	**COM1**
01 Valle dell'Arno	*2.5*	56.1	75.0	41.3	41.6	*0.4*	*17.3*	84.1	82.8	41.4	41.9
02 Lecchese	*3.2*	49.4	54.9	30.5	35.9	*2.3*	*69.6*	54.2	57.3	30.7	36.2
03 Valli Bresciane	*5.4*	56.3	56.0	38.4	43.6	*3.1*	*56.8*	59.8	58.4	38.6	44.0
04 Serico comasco	*2.7*	55.0	57.3	35.0	37.8	*1.1*	*41.2*	66.1	65.0	35.2	38.4
05 Valseriana	*1.4*	62.0	75.6	37.5	39.0	*0.5*	*37.3*	79.4	81.4	37.7	39.3
06 Castelgoffredo	*1.4*	60.4	66.6	17.6	19.4	*0.6*	*43.9*	70.6	72.7	18.1	19.9
07 Bassa Bresciana (cuoio calzature)	*1.3*	66.0	84.8	24.9	26.0	*0.4*	*28.7*	85.7	88.5	25.3	26.3
08 Sebino	*1.3*	67.5	84.9	38.7	40.4	*0.4*	*30.3*	84.1	88.4	38.8	40.6
09 Est Milanese	*11.8*	50.5	51.7	49.3	50.4	*3.0*	*25.6*	54.9	54.3	49.3	51.0
10 Brianza	*8.6*	44.8	44.1	48.8	46.3	*4.9*	*57.3*	48.1	46.3	48.9	47.0
11 Bergamasca-Val Cavallina-Oglio	*3.2*	72.3	77.1	25.4	28.7	*1.2*	*37.6*	76.4	79.6	25.8	29.0
12 Lecchese tessile	*0.6*	77.8	91.7	16.7	17.0	*0.1*	*25.7*	98.0	94.0	17.3	17.6
13 Bassa Bresciana (confezioni abbigliamento)	*1.8*	60.5	76.5	38.9	39.9	*0.5*	*26.5*	81.9	83.0	39.0	40.2
14 Gallaratese	*3.6*	52.0	51.3	51.7	50.8	*1.4*	*38.6*	60.4	57.5	51.8	52.1
15 Vigevanese	*2.7*	54.0	52.8	68.5	59.0	*1.0*	*38.7*	59.3	57.2	68.5	60.3
16 Casalasco Viadanese	*1.5*	47.8	58.8	33.4	36.5	*0.6*	*38.4*	62.9	66.5	33.6	37.0
Total Industrial districts		**58.3**	**66.2**	**37.3**	**38.3**			**70.4**	**70.8**	**37.5**	**38.8**
Pearson corr. coeff. with rapp_1		*-0.54*	*-0.65*								
Pearson corr. coeff. with rapp_2								*-0.79*	*-0.81*		
Pearson corr. coeff. with rapp_3								*-0.68*	*-0.64*		

The analysis of the results by industrial districts is much more controversial with respect to the unemployed counts. Table 6 shows for each district a selection of the results previously described for the whole set of districts. The results may be summarised as follows.

As previously noted, the small stock of unemployed in each industrial district leads to inefficient direct estimations due to the high sample errors in almost all the areas. In particular:

- HT and POS estimators are highly biased in all the industrial districts, showing less variable ARB percent values (from a minimum of 44.8 to a maximum of 77.8 for the HT estimator and from 44.1 to 91.7 for the POS estimator) with respect to the employed counts.

- SYS1 results the most efficient estimator in almost all districts and only in two districts (10 and 15) POS results less biased than SYS1. For these districts it is possible to gain efficiency introducing the composite estimator COM1, which shows a RRMSE lower than that of SYS1.
- As for the employment counts, the precision of the direct estimators seems to be negatively correlated to rapp_2 and, to a lesser extent, to rapp_3, signalling that, apart from the size of the stock under investigation, there is a clear negative association between the relative size of the district and the precision of direct estimations. In particular in the four larger districts where the ratio rapp_2 is higher than 2% (districts 2, 9, 3 and 10), the RRMSE of both HT and POS estimators is not higher than 60% . In addition, it is worth noting that these districts are characterized by the higher sample rates rapp_1 , with values of ARB under the mean of all districts.

5 CONCLUSIONS

The sampling rates in Lombardy's industrial districts are quite small and this evidence obviously affects the precision of the estimates.

In this context the estimates are more reliable for employment than for unemployment. This is essentially due to the different size of the stocks under investigation, the unemployed counts being significantly lower in Lombardy's industrial districts.

As regards employment, the HT estimator results, on the whole, the less biased if we exclude a district where the very high bias is significantly affected by the low sample rate. The POS estimator is, on average, more bias. This is essentially due to the ratio bias characterizing its formulation which is, in many cases, particularly high. In addition, the bias of the POS estimator presents a high cross-section variability which is, in turn, negatively correlated to the ratio rapp_1. This ratio indicates the weight expressed in terms of population of the district on the strata in which the district is located. Conversely, POS estimator is more efficient than HT, indicating that the post stratification by sex-age groups is able to increase the precision of the estimates. In addition, it is worth noting that the precision of both direct estimations seems to be negatively correlated to rapp_2 and, even if to a lesser extent, to rapp_3.

The different synthetic estimators introduced present low values of percent ARB and this means that Lombardy's districts do not exhibit strong individual effects with respect to the respective macro-area pattern of employment. The synthetic estimator with regional macro-area shows, on average, the higher efficiency. This is largely as expected, as this estimator is based on the largest set of sample units, thus reducing the variability of the estimates. In 6 districts out of 16 it is possible to gain higher efficiency, without increasing bias, by introducing alternative formulations of the synthetic estimator (e.g., SYN2 for the district coded by 4, 6 and 10, SYN3 for the districts coded by 1 and 14), thus signalling that specific composite estimators for each industrial district may be studied by applying different definitions of the reference macro-areas.

The adoption of the composite estimator generally does not increase the efficiency of the estimations. Only in three districts COM1 results the most efficient estimator (because of the lower bias of the post-stratified ratio estimator), while in the remaining district COM1 and SYS1 present the same percent RRMSE.

As regards unemployment, the results are more controversial. Both direct estimators are highly biased in every industrial district. The POS estimator presents the same values of the percent RRMSE as the HT estimator, thus signalling that, for the unemployment counts, the post-stratification based on demographic auxiliary variables does not play a role in improving the quality of the estimations. On another hand, as for the employment counts, it emerges a positive association between the ratio rapp_2 (and, even if to a lesser extent, rapp_3) and the precision of the direct estimates.

Finally, the synthetic estimator with regional macro area (SYS1) presents the higher efficiency in almost all districts. Only in two cases it is possible to improve efficiency by introducing the composite estimator COM1.

These results, although limited to a small group of small areas, point out the need for further research on small area estimation. In particular, with reference specifically to labour market stocks, more attention must be paid to unemployment, by introducing both model-based estimations, and by adopting more suitable auxiliary variables at the local area level.

REFERENCES

Cicchitelli G., Herzel A., Montanari G. (1992) *Il campionamento statistico*, Il Mulino, Bologna.

Cochran W.G. (1977) *Sampling Techniques*, Wiley, New York, 3rd edition.

Falorsi P.D. and Falorsi S. (1994) *Stime trimestrali a livello provinciale per l'indagine sulle Forze di Lavoro*, Quaderni di ricerca ISTAT, n.3,.

Faramondi A. and Piras M. G. (2002) *Le nuove stime di aggregati socio-economici per i Sistemi Locali del Lavoro*, Sviluppo Locale, 20, 80-95.

ISTAT (1991) *Forze di lavoro: disegno dell'indagine e analisi strutturali*, Annali di Statistica, Roma.

Rao J.N.K. (2003) *Small Area Estimation*, Wiley, New Jersey.