Measuring poverty and social exclusion by small area estimation

D'Alò M.[1], Bertarelli G.[2], Daddi S.[1],Di Consiglio L.[1], Fasulo A.[1], Ranalli M.G. [3], Solari F.[1], Casacci S.[1], Ercolani D.[1], Guandalini A.[1], Pichiorri T., [1]

Keywords: poverty rate, quantile share ratio, low work intensity, mixed models, M-quantile regression, latent Markov models.

1. Introduction

The objective of this work is to provide a statistical tool that can drive local policies on the basis of urban specificities. For this purpose, very detailed and updated statistical information at fine geographic level is necessary. Typically, the former aspect has always been assured by the Census that until now had the limit of providing data on a decennial basis. Such a temporal discrepancy is no longer acceptable nowadays. The timeliness of the information is, on the other side, assured by sample surveys, which, unfortunately, have limitations on the territorial level dissemination: the estimates are, in fact, usually produced at regional level. From these considerations, it emerges the need to provide solutions that exploit the availability of new sources of information, such as administrative data. The integration of this information with survey data can overcome the lack of information at a more detailed territorial level, assuring simultaneously timely and accurate estimates. NSIs have started to produce social and economic indicators using administrative data at local level. However, due to a different taxonomy, these indicators do not coincide with those usually computed by means of sample surveys. Therefore, the information from administrative data is often not consistent with the information officially produced at the regional level with sample surveys. The aim of this work is, first of all, to compare the indicators computed by the two sources of information, for all the metropolitan cities in Italy, for some large municipalities and for functional aggregations of small municipalities. The following step is to use the administrative data as an auxiliary source for model based estimation or for projection-type estimators. The output of this step allows us to evaluate the results obtained on important indicators of social exclusion and well-being, typically produced with the EuSilc (European Union Statistics on Income and Living Conditions) survey. In particular, we focus small area estimates of poverty rate, low work intensity and quantile share ratio indicators, computed at provincial and metropolitan municipalities level.

2. Methods

The indicators of social exclusion and well-being are usually estimated at regional level (NUTS2) with the EuSilc data, that is planned to be reliable at that level. However estimates at a more disaggregated level are in many cases too inaccurate. Nowadays, administrative data can be used to improve estimation either directly or as a source of auxiliary variables in model based estimation.

[1] Istituto nazionale di Statistica

[2] Università di Pisa

[3] Università di Perugia

In this work we aim at comparing the direct estimates with the administrative source based indicators and the small area estimators. The indicators derived from administrative data take usually larger values than those estimated by means of survey data: this provides evidence that the indicators derived from administrative data are not consistent with the survey statistics. This is due to several factors such as differences of definition of income items between the two sources; errors in the quantification of income in administrative sources - the undercoverage of some income categories can be greater in some areas; over or undercoverage of specific populations - e.g. foreigners.

However, thanks to the strong correlation between the survey variables and the proxy variables from administrative data, these can be very useful for small area modelling. Different types of models can be fitted and a decision needs to be taken on which type of model is best suited to the available data. The small area applications are based on two types of models that can be viewed as special cases of the general mixed linear model. If the data involved into the general modelling refer to small areas of interest an area-type model (see Fay Herriot, 1979) is fitted, whereas a basic unit model (see Battese et al., 1988) is defined when these data refer to the units belonging to the small area of interest. The auxiliary variables used in these models include the "equivalent" administrative indicator variables of the social exclusion status for each of the target indicator. In particular, we consider the following target indicators: at risk of poverty rate, low work intensity, income inequality.

In the small area model the auxiliary information is at small area level. In order to boost the predictive capacity of the model, the proxy of the target indicators computed with the administrative data are also considered into the model. Other variables included are: the percentage of foreigners, employment rate, population distribution by gender and age classes.

In our analysis we considered two sources of data. Survey Indicators are computed using EuSilc 2016 (reference year 2015) and the administrative data comes from the statistical register ARCHIMEDE 2015, an integrated archive of socio-economic and demographic microdata produced by ISTAT (Garofalo, 2014; Wallgren and Wallgren, 2007). The unplanned domains of interest are all the 14 metropolitan cities in Italy and the 110 Italian provincial areas (excluding metropolitan cities), for a total of 124 small areas.

3. RESULTS

Figure 1 shows the empirical distribution, pairwise scatterplots, and correlation at domain level between the direct estimates and the corresponding indicators computed with administrative data.

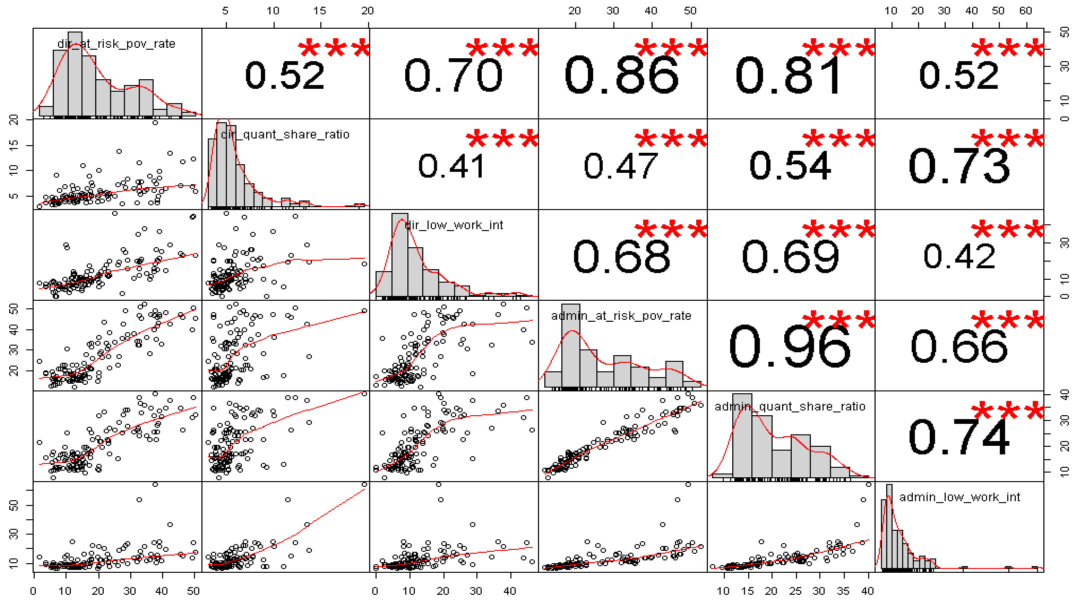

Figure 1. Correlation between direct estimates and correspondent proxy computed with administrative data. From the first to the last row: at risk of poverty rate, low work intensity, income inequality from EuSilc and at risk of poverty rate, low work intensity, income inequality from the statistical register- reference year 2015.

The links between survey data and administrative data allow the use of indicator variables taken from administrative archives and also used in this case to define unit-level models with greater predictive power. Figures 2 and 3 show the estimates and corresponding CV of the direct, administrative indicators and area level small area estimators for the at risk of poverty rate and the low work intensity, respectively.

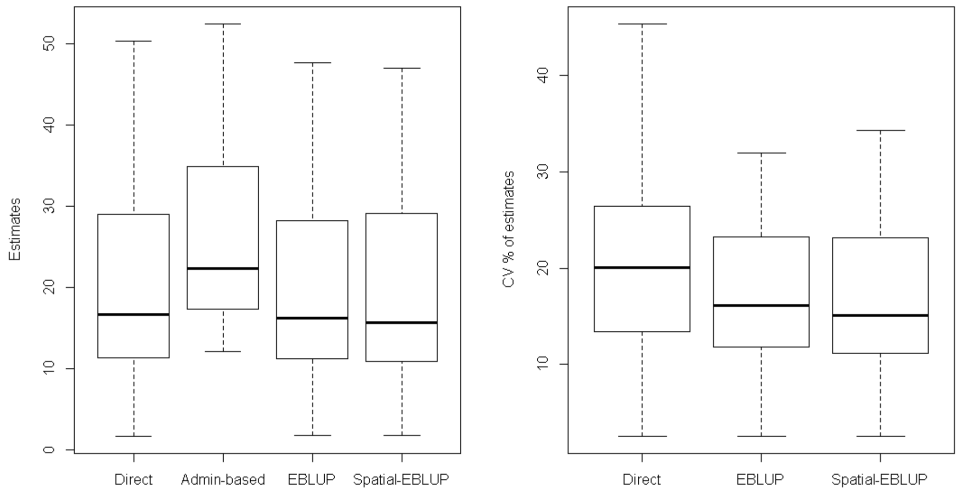

Figure 2. Distribution of at risk poverty rate estimates and the correspondent CV %

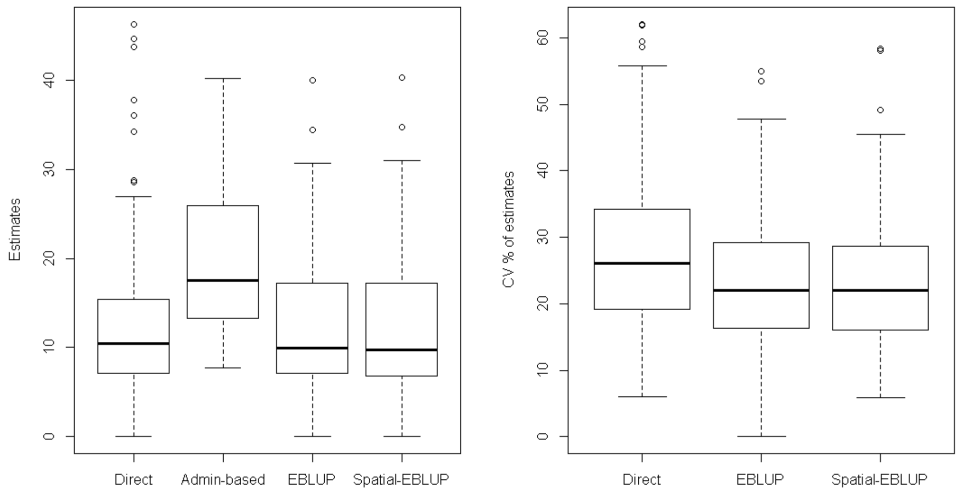

Figure 3. Distribution of low work intensity estimates and the correspondent CV %

The values of the indicators computed with the administrative source are generally higher than those obtained with EuSilc. This effect can be due to several factors, also concomitant, such as (i) differences in the definition of the income components between the two sources – gross income in the administrative source, net income in EuSilc – , (ii) differences in the population – household in EuSilc and individuals in the administrative source – (iii) errors in the quantification of income in administrative sources, and (iv) undercoverage of some income categories.

The use of administrative data in the small area model allows the improvement the direct estimates in terms of accuracy without a relevant effect on bias. Similar results, not shown here, are observed for unit level models. Moreover, spatial correlation among areas can be introduced both in area and unit level model (Pratesi and Salvati, 2009, Saei and Chambers 2009). The results show that incorporating the spatial structure provides larger gains in efficiency, in particular for the smallest areas. Finally, in particular for the quantile share ratio, the M-quantile estimator (Chambers and Tzavidis, 2006) may be more effective to estimate the distribution function and therefore the quintiles used for the definition of the of indicator. Also, the approach that uses latent Markov models (Bertarelli et al., 2018) can be used to exploit the multivariate nature of the problem and correlation structure among the indicators.

References

[1] Fay, R. E. and R. A. Herriot (1979). Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association 74, 269277.

[2] Battese, G., Harter, R., & Fuller, W. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. Journal of the American Statistical Association, 83(401), 28-36. doi:10.2307/2288915

[3] Garofalo, G. (2014). Il progetto ARCHIMEDE: obiettivi e risultati sperimentali. Technical report, ISTAT, Working papers, 9. Available at: https://www.istat.it/it/files//2014/11/IWP-n.-9-2014.pdf

[4] Wallgren, A., & Wallgren, B. (2007). Register-based statistics: administrative data for statistical purposes (Vol. 553). John Wiley & Sons

[5] Pratesi, M., N. Salvati, et al. (2009). Small area estimation in the presence of correlated random area eects. Journal of O-cial Statistics 25(1), 37.

[6] Saei, Ayoub and Chambers, Ray (2003). Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects Southampton, UK. Southampton Statistical Sciences Research Institute 31pp. (S3RI Methodology Working Papers, M03/15)

[7] Chambers, R., Tzavidis, N. (2006) M-quantile models for small area estimation. Biometrika 93:255–268

[8] Bertarelli, G., Ranalli, M. G., Bartolucci, F., D'Alò, M. & Solari, F. (2018). Small Area Estimation for Unemployment using latent Markov models. Survey Methodology, 44(2) (in print)