Rivista Italiana di Economia Demografia e Statistica Volume LXVIII n. 3/4 Luglio-Dicembre 2014

A COMPARISON OF BIAS CORRECTION METHODS FOR THE DISSIMILARITY INDEX

Anna Maria Altavilla, Angelo Mazza, Antonio Punzo

1. Introduction

The segregation of demographic groups, often connected to ethnicity, age or gender, is an important area of research among sociologists, demographers and other social scientists. The evaluation of segregation within a population is typically based on the proportions of demographic groups belonging to some kind of allocation units, such as residential areas, workplaces, or schools (Mazza and Punzo, in press).

Many segregation indexes have been suggested, with different formulations denoting different definitions of segregation (see Massey and Denton, 1988 for an overview). Among these, the dissimilarity index *D*, proposed by Duncan and Duncan (1955), is widely used to assess the differential distribution of two groups among allocation units. This index has been used in a broad range of contexts, such as gender segregation (see, e.g., Karmel and Maclachlan, 1988), labor force segregation (for a survey see Flückiger and Silber, 1999), and residential segregation (see Duncan and Duncan, 1955, and Massey and Denton, 1987, 1988).

Generally, the observed settlement pattern is the resultant of a mix of behavior-based forces; thus it should be seen as one of the many possible outcomes of a stochastic - rather than deterministic - allocation. Usually researchers are interested in understanding the "systematic" characteristics of the allocation process, apart from random fluctuations that may affect a single observed pattern (Altavilla, Mazza, Punzo, 2012). In this view, the observed dissimilarity $\widehat{D}$ is merely an estimator of a true but unknown level of dissimilarity in the population D. So, it should be clear why this randomness also holds even if the index is computed on a full-count census data. A problem with the use of this index is that $\widehat{D}$ appears to be an upward biased estimator of D. Within a multinomial framework based on the assumption that individuals allocate themselves independently and that unit sizes are not fixed (see Section 2), Allen et al. (2009) demonstrate, using simulations, that random allocation generates substantial unevenness, and hence an upward bias, especially when dealing with small units, a small minority proportion, and a low level of seg-

regation. Accordingly, different correction approaches have been proposed in literature (see, e.g., Allen et al., 2009, and Altavilla, Mazza and Punzo, 2010 for two examples of bootstrap-based bias correction, Altavilla, Mazza, Punzo, 2012 for an analytical computation of bias and Mazza and Punzo 2014 for a new bias correction which outperforms all previous correction attempts).

In the following, four bias correction techniques, based on grouped jackknife, bootstrap, double bootstrap and the Mazza and Punzo (in press) proposal, are compared in terms of their mean bias. The paper is organized as follows. In section 2, inferential framework and notation are given, in section 3 the four estimators are described and in section 4 there is their comparison. Finally, in section 5, conclusions are drawn.

2. Inferential framework and notation

Consider an area subdivided into k subareas (or units), denoted by $j = 1, \ldots, k$, being populated by n individuals according to a dichotomous characteristic indexed by $c = 0,1$. Examples of common dichotomous characteristics are black or white ethnicity, male or female gender, and so on. The number of individuals with status c is denoted by n^c, $c = 0,1$, with $n = n^0 + n^1$. There will be n_j^c individuals in unit j having status c, with $n^c = \sum_{j=1}^{k} n_j^c$, $c = 0,1$. The observed settlement -- characterized by the two sets denoted by $n_1^0, \ldots, n_k^0$ and $n_1^1, \ldots, n_k^1$ -- is, however, only one of the possible realizations of an underlying *allocation process* P. If it is plausible to assume that individuals allocate themselves independently and that unit sizes are not fixed, then the process will be governed by the conditional probabilities

$$p_j^c = P(unit\ of\ membership = j|c), \qquad j = 1, \ldots, k, \qquad c = 0,1 \qquad (1)$$

that an individual i will belong to the unit j, given his/her status c.

Social scientists are usually interested in making inferences on a particular function of these probabilities; this function, commonly called "segregation index", should express the degree of segregation that characterize the process P. Before to introduce any kind of segregation index, it is important to define the concept of systematic segregation, occurring when there is at least one subarea in which individuals belonging to the two groups have a different probability to allocate themselves; in mathematical terms this means that:

$\exists\ j:\ p_j^1 \neq p_j^0.$

Among the many segregation indexes existing in literaturem, the most popular one is without doubt the Duncan and Duncan (1955) segregation index, usually denoted by D, characterized by the formula:

$$D = \frac{1}{2}\sum_{j=1}^{k} \left| p_j^1 - p_j^0 \right| \tag{2}$$

Obviously, the index in (2) takes values on the compact interval $[0,1]$ and it increases as systematic segregation grows. Furthermore, it is straightforward to note that the case $D = 0$ (absence of systematic segregation) is achievable if, and only if

$p_j^1 = p_j^0 \quad \forall\, j$.

Unfortunately, we can only observe the crude counterpart of D

$$\widehat{D} = \frac{1}{2}\sum_{j=1}^{k} \left| \frac{N_j^1}{n^1} - \frac{N_j^0}{n^0} \right| = \frac{1}{2}\sum_{j=1}^{k} \left| \hat{p}_j^1 - \hat{p}_j^0 \right| \tag{3}$$

where $\hat{p}_j^c$, proportion of individuals with status c in the unit j, $c = 0,1$, is the plug-in estimator of p_j^c. The word "unfortunately" is justified if one thinks that the observed settlement pattern is only one of the numerous possible patterns arising from P, each of them with probability (see Allen *et al*., 2009) given by the product of two independent multinomial distributions, one for $c = 0$ and one for $c = 1$:

$$P(n_1^c, \ldots, n_k^c \mid p_1^c, \ldots, p_k^c, n^c) = \prod_{j=1}^{k} \prod_{c=0}^{1} n^c! \frac{\left(p_j^c\right)^{n_j^c}}{n_j^c!} \tag{4}$$

3. Estimators

In this section, we introduce four alternative bias correction techniques.

1.1. Bootstrap based estimator

With the aim to eliminate, or at least reduce, the upward bias of $\widehat{D}$, Allen et al. (2009) adopt a bootstrap-based bias correction. It is based on the idea that

$$D - \widehat{D}_{\text{obs}} \approx \widehat{D}_{\text{obs}} - E\left(\widehat{D}|\hat{p}_1^0, \dots, \hat{p}_k^0, \hat{p}_1^1, \dots, \hat{p}_k^1, n^0, n^1\right), \quad (5)$$

where $\widehat{D}_{\text{obs}}$ denotes the observed counterpart of $\widehat{D}$. The observed conditional probabilities $\hat{p}_j^0$ and $\hat{p}_j^1$, $j = 1, \dots, k$, are used to generate, by multinomial sampling, B bootstrap allocations with the same group sizes n^0 and n^1. Then, a measure of $\text{Bias}\left(\widehat{D}\right)$ is given by $\overline{D}_{\text{Boot}} - \widehat{D}$, and the bootstrap bias corrected estimate of D can be obtained as

$$\widehat{D}_{\text{Boot}} = \widehat{D}_{\text{obs}} - \left(\overline{D}_{\text{Boot}} - \widehat{D}_{\text{obs}}\right) = 2\widehat{D}_{\text{obs}} - \overline{D}_{\text{Boot}}. \quad (6)$$

This bias correction would work well if the bias were constant for different values of D. This is not the case here, and this bias correction is therefore not expected to "eliminate", but only to "reduce", the existing bias. Instead of bootstrapping $E\left(\widehat{D}|\hat{p}_1^0, \dots, \hat{p}_k^0, \hat{p}_1^1, \dots, \hat{p}_k^1, n^0, n^1\right)$, Mazza and Punzo (in press) show that this expectation may be computed analytically, using a binomial based formulation for a small number of units with small sizes or with a folded normal approximation when n^c, $c = 0{,}1$, is sufficiently large.

1.2. Grouped jackknife and iterative bootstrap estimators

Alternative to the bootstrap, a standard practice for bias correction is the Jackknife. Hence, we evaluated, also, a grouped jackknife estimator $\widehat{D}_{\text{JK}}$; this estimator has been implemented following Efron (1982, Section. 2.2). Finally, a double bootstrap estimator $\widehat{D}_{\text{DB}}$, based on the approach documented in Davison and Hinkley (1997, Section. 3.9) has also been evaluated.

1.3. A recently introduced estimator

Mazza and Punzo (2014) introduce a new estimator of D, which further reduces the bias with respect to $\widehat{D}_{\text{Boot}}$. Its rationale consists in choosing a value $\widetilde{D}$ which minimizes

$$E\left(\widehat{D}|\tilde{p}_1^0, \dots, \tilde{p}_k^0, \tilde{p}_1^1, \dots, \tilde{p}_k^1, n^0, n^1\right) - \widehat{D}_{\text{obs}} \quad (7)$$

with $\widetilde{D} = \frac{1}{2}\sum_{j=1}^{k} \left|\tilde{p}_j^1 - \tilde{p}_j^0\right|$. There may be different criteria for choosing $\widetilde{D}$. On way is to require the sequence of differences $\left|\tilde{p}_j^0 - \tilde{p}_j^1\right|$ to be a flattened variant of its observed counterpart. Flattening is obtained by spreading the difference $\Delta = \widehat{D}_{\text{obs}} - \widetilde{D} \geq 0$, among the k differences $\left|\tilde{p}_j^0 - \tilde{p}_j^1\right|$, proportionally to the residuals $\hat{d}_j = \left|\hat{p}_j^0 - \hat{p}_j^1\right|$. An optimization procedure, which adopts a combination of golden section search and successive parabolic interpolation is described in Mazza and Punzo (2014).

4. Comparison of estimators

In this section we use Monte Carlo simulations to compare the bias of $\widehat{D}$, and of the four estimators $\widehat{D}_{\text{Boot}}$, $\widehat{D}_{\text{JK}}$, $\widehat{D}_{\text{DB}}$ and , $\widetilde{D}$ described in the previous section. The setup of the simulations is similar to the one adopted by Allen et al. (2009). The sets of conditional probabilities $p_1^0,\ldots,p_k^0$ and $p_1^1,\ldots,p_k^1$, with $k=50$, were obtained with the formula

$$P(\text{unit} \le j \mid c=1) = \frac{(1-q)P(\text{unit} \le j \mid c=0)}{1-qP(\text{unit} \le j \mid c=0)} \tag{8}$$

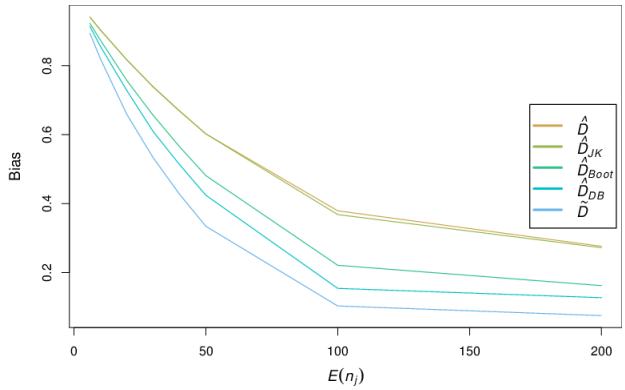
proposed in Duncan and Duncan (1955); it may be observed that each value of q is related to one value of D. Although this set of segregation curves cannot represent all distributions of segregation, it is a sufficient set to examine different levels of systematic segregation for the purposes of this paper. The formula above, combined with the constraint of equal expected unit sizes $E(n_j)$, fixes the conditional allocation probabilities for both groups. An allocation is then generated by assigning n^1 and n^0 individuals to the k units by sampling from two multinomial distributions having each one of the two sets of conditional probabilities as parameter.

The simulation factors considered are p, $E(n_j)$ and D. For each of them, a grid of values is chosen: 0.01, 0.05, 0.1, 0.3, and 0.5 for p; 6, 10, 20, 30, 40, 50, 100 and 200 for $E(n_j)$; 0, 0.056, 0.127, 0.225, 0.292, 0.382, 0.634, and 0.818 for D. Values chosen for D are respectively related to the values 0, 0.2, 0.4, 0.6, 0.7, 0.8, 0.95, and 0.99, of the parameter q in the previous equation. The number of units is fixed at $k = 50$ and the number of bootstrap replications is fixed to B = 100. For each combination of the considered simulation factors, 1000 samples are generated randomly.

The mean simulated biases of the estimators considered are depicted in the figures below.

It may be noted that when p, $E(n_j)$ and D present low values, the bias of $\widehat{D}$, the uncorrected estimator, is considerably high, incorrectly suggesting that a highly segregating process underlies the allocation. In the opposite situation of high values of p, $E(n_j)$ and D, all estimators provide values not very different from the true value D. From these results, we can note as $\widetilde{D}$ most often outperforms all other estimators in reducing the bias, while the grouped jackknife estimator, in all the considered scenarios of simulations, showed only a negligible improvement over $\widetilde{D}$.

Figure 1 – *Comparison between biases at the varying of* $\boldsymbol{E(n_j)}$*, fixed p = 0.01 and* $\boldsymbol{D}$*=0 .*



As to the double bootstrap approach, the added level of bootstrap did improve the performance of in terms of mean bias over $\widehat{D}_{\mathrm{Boot}}$; however, these improvements were only marginal, and very far from counterbalancing the higher computational burden required.

Figure 2 – *Comparison between biases at the varying of D, fixed p = 0.01 and* $\boldsymbol{E(n_j)}$ *= 20.*

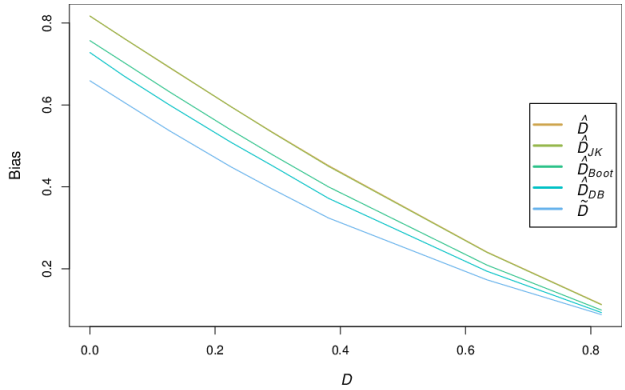
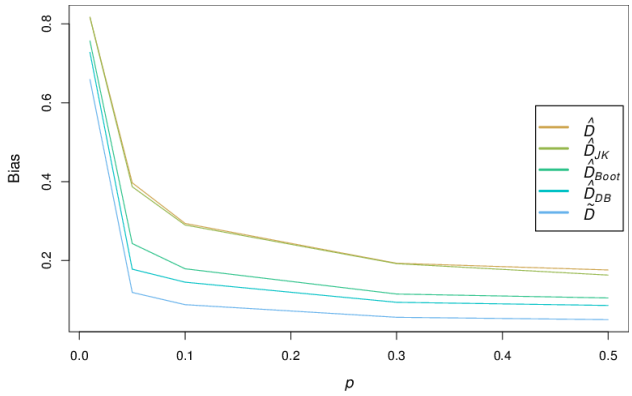


Figure 3 – *Comparison between biases at the varying of p, fixed D = 0 and* $\boldsymbol{E(n_j)}$ *= 20.*



5. Conclusions

It has long been recognized that the sensitivity of the dissimilarity index of Duncan and Duncan (1955) to random allocation implies an upward bias, particularly evident with smaller unit sizes, small minority proportions and lower levels of segregation. In this paper, following a multinomial framework, we have compared, using Monte Carlo simulations, the performance of four bias reduction techniques, based on bootstrap, grouped jackknife, double bootstrap and on a recent procedure introduced in Mazza and Punzo (2014). This new procedure performed better than its competitors did, although for reliable estimations, minority proportion and unit sizes do not have to be both very small. The grouped jackknife bias-corrected estimator exhibited only a little improvement over the natural estimator and so did the double bootstrap estimator with respect to the bootstrap bias-corrected one.

References

ALLEN, R., BURGESS, S., WINDMEIJER, F. (2009). More reliable inference for segregation indices. *Technical Report 216*, The Centre for Market and Public Organisation, University of Bristol.

ALTAVILLA A.M., MAZZA A., PUNZO A. (2010). Sull'impiego di un indice di dissimilarità nello studio della disposizione di popolazioni straniere su un territorio urbano. *Rivista Italiana di Economia, Demografia e Statistica*, vol. LXIV, p. 7-14.

ALTAVILLA A.M., MAZZA A., PUNZO A. (2012). On the upward bias of the dissimilarity index. *Rivista Italiana di Economia, Demografia e Statistica*, vol. LXVI – N. 1, p. 15-20.

DAVISON, A. C., HINKLEY, D. V. (1997). Bootstrap Methods and Their Application, volume 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

DUNCAN, O. D., DUNCAN, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, 20(2), 210–217.

EFRON, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

FLÜCKIGER, Y., SILBER, J. G. (1999). *The measurement of segregation in the labor force*. Physica-Verlag, Heidelberg.

KARMEL T., MACLACHLAN M. (1988). Occupational sex segregation - increasing or decreasing? *Economic Record*, 64(3), 187–195.

MASSEY D. S., DENTON, N. A. (1988). The dimensions of residential segregation. *Social Forces*, 67(2), 281–315.

MAZZA A., PUNZO A. (in press). On the upward bias of the dissimilarity index and its corrections. *Sociological Methods & Research*.

SUMMARY

The dissimilarity index of Duncan and Duncan is widely used in a broad range of contexts to assess the overall extent of segregation in the allocation of two groups in two or more units. Its sensitivity to random allocation implies an upward bias with respect to the unknown amount of systematic segregation. In this paper, following a multinomial framework based on the assumption that individuals allocate themselves independently and that unit sizes are not fixed, we report the results of Monte Carlo simulations performed in order to compare the natural estimator with four bias reduction techniques, based on bootstrap, grouped jackknife, double bootstrap and on a more recent procedure. Results indicate the new procedure performed better than its competitors did, although for reliable estimations, minority proportion and unit sizes do not have to be both very small.

Anna Maria ALTAVILLA, University of Catania, Department of Economics and Business, altavil@unict.it

Angelo MAZZA, University of Catania, Department of Economics and Business, a.mazza@unict.it

Antonio PUNZO, University of Catania, Department of Economics and Business, antonio.punzo@unict.it