VOLUME LXXI – N. 2 APRILE-GIUGNO 2017

RIVISTA ITALIANA DI ECONOMIA DEMOGRAFIA E STATISTICA

Sede Legale: C/O Studio Associato Cadoni, Via Ravenna n.34 – 00161 ROMA.
sieds.new@gmail.com, rivista.sieds@gmail.com

SIEDS
SOCIETÀ ITALIANA
DI ECONOMIA DEMOGRAFIA E STATISTICA

Sede Legale:
C/O Studio Associato Cadoni, Via Ravenna n.34 – 00161 ROMA

sieds.new@gmail.com

rivista.sieds@gmail.com

INDICE

DEALING WITH THE BIAS OF THE DISSIMILARITY INDEX OF SEGREGATION[1]

Angelo Mazza

1. Introduction

Ethnic residential segregation has long been investigated, especially in the USA and South Africa. A series of papers, dating back to the late 1940s and early 1950s and mostly published in the American Sociological Review, address the issue of measuring segregation and introduce a wide variety of indices (e.g.: Wright (1937), Jahn, Schmid, and Schrag (1947), Williams (1948), Jahn (1950), Cowgill and Cowgill (1951), Bell 1954).

In 1955 in a celebrated article Duncan and Duncan provided a systematic analysis and critique of these segregation indexes, showing that all of them could be regarded as functions of a single geometrical construct, the "segregation curve". This implies that these indexes are related and have mathematical properties that often lead to difficulties of interpretation. This brought Duncan and Duncan (1955) to assert “the status of the empirical work already done with segregation indexes is questionable, and their validity for further research is undetermined”. The authors also proved that most of the previously proposed statistics were mathematically related to the dissimilarity index *D* (see next paragraph for its definition).

For nearly twenty years, *D* served as the undisputed segregation measure, routinely employed to measure segregation between social groups. The consensus among a vast variety of scholars, comprising geographers, sociologists and even some economists, was so unanimous that this period has been described by Massey and Denton (1988) as the “Pax Duncana”.

In 1976 the Pax Duncana came abruptly to an end, with the publication on the American Sociological Review of a critique of the dissimilarity index by Cortese, Falk, and Cohen. The major objection of Cortese and his coauthors to *D* is that it postulates the expectation of evenness as the opposite of segregation, whereas in most cases it is not as useful as the concept of randomness.

In fact, *D* takes its minimum value of zero only under the condition of exact even distribution, which usually is not possible because individuals, families, and

[1] Invited paper to the 54th SIDES Scientific Meeting – Catania 2017.

households cannot be distributed in fractional parts, as it is often needed to achieve exact even distribution (Taeuber and Taeuber 1965; Fosset, 2017). As written by Voas (2000), "even a completely random distribution of individuals could produce areas of an uneven composition, just as a well-shuffled deck of cards will still produce hands with unbalanced suits". At the opposite, within a model of random distribution, in which race and neighborhood are statistically independent, an exact even distribution is a highly unexpected outcome and its occurrence can signal that race is 'systematically' associated with residence through some kind of structured social dynamic, such as a group quota allocation process (Fosset, 2017).

Because of these factors, *D* is inherently subject to an upward bias. The random effects are not great when the sets concerned are large, but if the area population is small or group proportion is very low, the index can be highly misleading. As a result, the fact that *D* is affected by differences in the proportion of the minority in the population and by the size of the areal unit of analysis (number of households), can result in misleading assessments of the level of segregation, and makes problematic intercity comparison, including the same city at another point in time.

It is worth noting that the above critics apply to other indices of segregation, like the Gini index, the Atkinson index, the Hutchens square root index, and the Theil entropy index.

In the following, we will give more details on the issue of the index bias and we will outline the main paths followed in literature to deal with this problem.

2. Inferential framework and notation

Consider an area subdivided into k spatial units denoted by $j = 1, \ldots, k$, and populated by n individuals characterised by a dichotomous attribute *c*, with *c*=0,1, such as black or white, male or female sex, and so on. The number of individuals with status c is denoted by n^c, with $n = n^0 + n^1$. There will be n_j^c individuals in unit j having status c, with $n^c = \sum_{j=1}^{k} n_j^c$.

It is important to note that the settlement observed is just 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

$$p_j^c = P(unit\ of\ membership = j|c), \qquad j = 1, \dots, k, \qquad c = 0,1 \tag{1}$$

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.

There is *systematic segregation* when there is at least one spatial unit where individuals of the two groups have a different probability to allocate themselves, i.e.:

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

Among the many segregation indexes proposed, the most popular is the dissimilarity index

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

Dissimilarity measures the percentage of a group’s population that would have to change residence for each neighborhood or area to have the same percentage of that group as the metropolitan area overall. The index ranges from 0 (absence of systematic segregation) to 1 (complete systematic segregation) and $D = 0$ if, and only if

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

However, in real life application we only know 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 $\hat{p}_j^c$ is the plug-in estimator of p_j^c. This is beacuse the observed settlement pattern is only one of the numerous possible patterns arising from P, each of them with probability given by the product of two independent multinomial distributions, one for $c = 0$ and one for $c = 1$ (see Allen *et al.*, 2009):

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

In this view, the observed dissimilarity $\hat{D}$ is merely an estimator of a true but unknown level of dissimilarity in the population D. Therefore, it should be clear why this randomness also holds even if the index is computed on a full-count census data (Altavilla, Mazza, Punzo, 2012).

3. Bias of of the dissimilarity index

We mentioned earlier that the value of the dissimilarity index $\hat{D}$ computed over the data observed is an estimator of the unknown systematic segregation D, so we can define its bias as

$$Bias\left(\hat{D}\right) = E\left(\hat{D}\right) - D. \tag{1}$$

The expectation in (5) can be explicited as follows:

$$E\left(\hat{D}\right) = \frac{1}{2} \sum_{\left(n_1^0,\dots,n_k^0\right):n^0} \sum_{\left(n_1^1,\dots,n_k^1\right):n^1} \left[\left(\sum_{j=1}^{k} \left| \frac{n_j^1}{n^1} - \frac{n_j^0}{n^0} \right| \right) \prod_{j=1}^{k} \prod_{c=0}^{1} n^c! \frac{\left(p_j^c\right)^{n_j^c}}{n_j^c!} \right] \tag{2}$$

where the first two summations run across all possible patterns $n_1^c, \dots, n_k^c$ satisfying the constraint $\sum_j n_j^c = n^c$.

It is important to note that $Bias\left(\hat{D}\right)$ tends to be positive, that is, $\hat{D}$ tends to overestimate the systematic segregation D. This is due to the fact that the index is based on absolute values; an intuitive explanation is that if, for instance, systematic segregation were 0, any sampling variation would result in an upward bias increase.

The problem of the index bias has received regular attention for over forty years; an early discussion is in Taeuber and Taeuber (1965), later Cortese et al (1976) and Winship (1977) addressed the issue.

Carrington and Troske (1997), Ransom (2000) and Allen et al. (2009) proved, using simulations, the nonnegativity of the bias, and it has been shown how it increases when dealing with small unit sizes, a small minority proportion, and a low level of segregation. Figure 1 shows the behaviuor of the bias as a function of minority proportion *p*, average size of spatial units $E(n_j)$, and systemic dissimilarity *D*.

Eventually, a mathematical proof of the nonnegativity of the bias was provided in Mazza and Punzo (2015).

4. Current practices for dealing with the index bias

Most researchers are aware that the index bias can have substantial impacts on results. However, because direct solutions to this problem have not been available, scholars most often rely on practices that Fosset (2017) defined as informal "rules of thumb", such as:

- restricting segregation studies to comparisons involving large and relatively balanced population groups;
- use larger spatial units such as census tracts;
- assess segregation using full count data;
- weight cases differentially, lowering cases presumed to be distorted by bias, when assessing variation in segregation over time or across groupings of cases.

Fosset (2017) argued that even if these practices for dealing with the bias did work, they still would have the undesirable consequence of restricting the scope of segregation studies. Furthermore, they preclude analyses at finer spatial scales, or involving populations small in absolute size. For instance, due to the concerns about index bias, studies assessing segregation at smaller spatial scales, once common, are currently very rare in literature.

More rigorous techniques for assessing the potential bias do have been proposed in literature. Analytic formulas have been proposed by Winship (1977), whereas Carrington and Troske (1997), Allen *et al.* (2009), and Mazza and Punzo (2015) proposed bootstrap-based methods. However, eliminating problematic cases, even if in a more formal way, would still narrow the scope of segregation studies.

5. Bias correction

If unbiased index scores were available, there would not be any need to exclude cases due to concerns about bias. Therefore, in time, several solutions for dealing directly with the index bias have been proposed, among others by Cortese et al. (1976), Winship (1977), Carrington and Troske (1997), Allen et al. (2009), and Mazza and Punzo (2015).

The bias reduction methods proposed are based on adjusting the index scores downward, in order to eliminate the effect of upward bias due to random allocation. These methods are based on the estimation of the expectation in (5), which requires the sampling distribution of *D*. Note that, as shown in Altavilla et al. (2009), being the bias a function of the systemic level of segregation, these corrections are only able to reduce the bias but not to eliminate it.

Mazza and Punzo (2015) showed that previous bootstrap-based bias correction could be obtained analytically, so without resorting to time-consuming resampling techniques. Mazza and Punzo also propose a new estimator; its rationale consists in choosing a value $\widetilde{D}$ that minimizes

$$E\left(\widehat{D}|\tilde{p}_1^0, \dots, \tilde{p}_k^0, \tilde{p}_1^1, \dots, \tilde{p}_k^1, n^0, n^1\right) - \widehat{D} \qquad (2)$$

with $\widetilde{D} = \frac{1}{2}\sum_{j=1}^{k} \left|\tilde{p}_j^1 - \tilde{p}_j^0\right|$.

Obviously, there may be different criteria for choosing $\widetilde{D}$; the authors chose to constrain the differences $\left|\tilde{p}_j^0 - \tilde{p}_j^1\right|$ to a flattened variant of their observed counterpart. Flattening is obtained by spreading the difference $\Delta = \widehat{D} \quad - \widetilde{D} \geq 0$, 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 (2015).

In Altavilla, Mazza and Punzo (2014), following a multinomial framework, the performance of four bias reduction techniques, based on bootstrap, grouped jackknife, double bootstrap and the procedure of Mazza and Punzo (2015) have been compared using Monte Carlo simulations. The procedure of Mazza and Punzo performed better than its competitors did, although for reliable estimations, minority proportion and unit sizes did not have to be both very small.

Most of the methods proposed do this using computation-intensive techniques. Fosset (2017) observed that these options introduce complexity and substantial computational burdens and so are unlikely to be widely adopted by researchers. This is indeed true; especially the lack of user-friendly computer programs has strongly affected a wider adoption of these proposals.

6. Conclusions

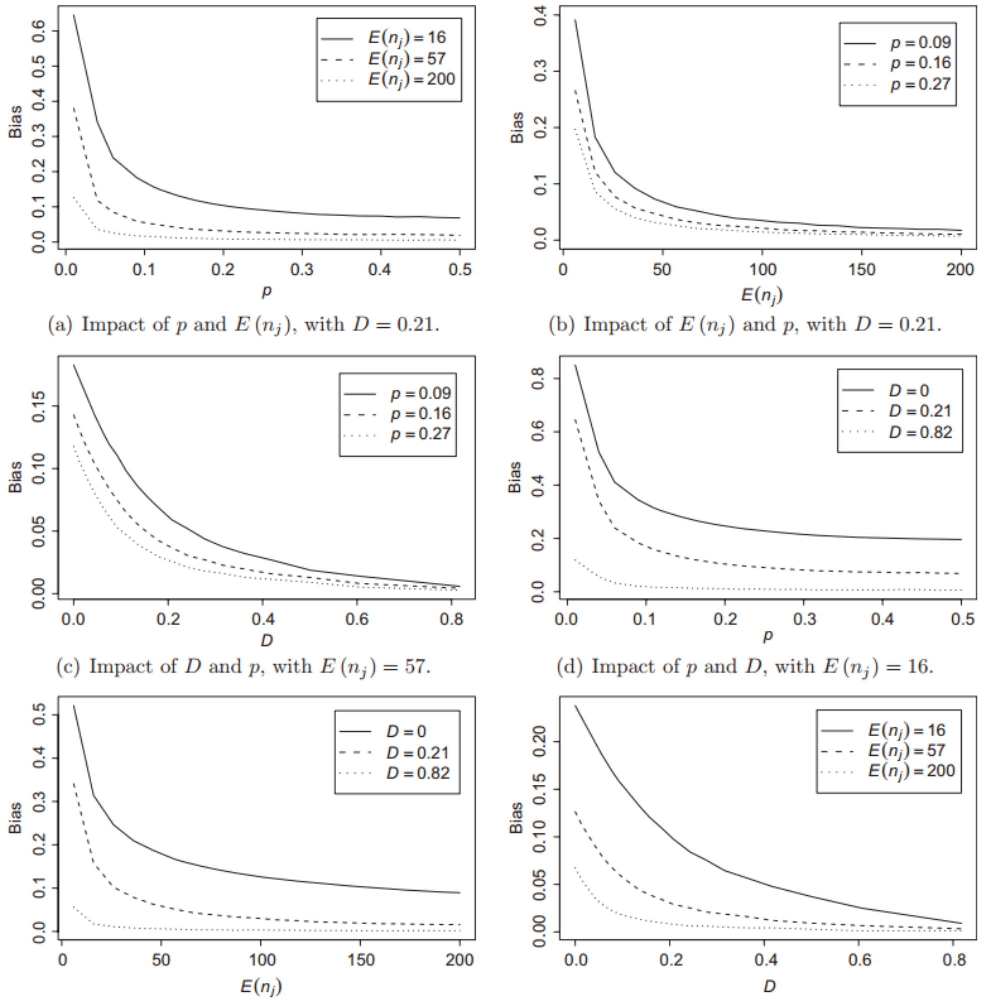
There is a widespread awareness that the index of dissimilarity, as much as other indices based on the segregation curve, are inherently subject to an upward bias. The index bias may not be relevant when the sets concerned are large, but if the area population is small or the minority proportion is very low, the index can be highly misleading.

Common strategies used in literature to deal with the index bias rely on the use of informal rules of thumb, which at least have the side effect of restricting the scope of segregation studies.

In time, solutions for dealing directly with the index bias have been proposed. They are based on adjusting the index scores downward, and require the computation of the sampling distribution of the index. Most of the methods proposed use computation-intensive techniques that have the drawback of introducing complexity and substantial computational burdens. The lack of user-friendly computer programs implementing these methods has strongly affected their wider adoption.

Appendix

Figure 1 – *Bias of the index of dissimilarity as a function of p, E(nj), and D.*



(a) Impact of p and $E(n_j)$, with $D = 0.21$.

(b) Impact of $E(n_j)$ and p, with $D = 0.21$.

(c) Impact of D and p, with $E(n_j) = 57$.

(d) Impact of p and D, with $E(n_j) = 16$.

(e) Impact of $E(n_j)$ and D, with $p = 0.11$.

(f) Impact of D and $E(n_j)$, with $p = 0.22$.

Results are obtained by Monte Carlo simulations using the parabolic segregation curves of Duncan and Duncan (1955) to generate the conditional probabilities (Mazza and Punzo, 2015)

References

ALLEN R., BURGESS S. and WINDMEIJER F., 2009. *More Reliable Inference for Segregation Indices*. The Centre for Market and Public Organisation, University of Bristol.

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, No. 1, pp. 15-20.

BELL W., 1954. A Probability Model for the Measurement of Ecological Segregation, *Social Forces,* Vol. 43, pp. 357– 64.

CARRINGTON W. J., TROSKE, K. R., 1997. On measuring segregation in samples with small units, *Journal of Business & Economic Statistics*, Vol. *15,* No. 4, pp. 402-409.

CORTESE C., FALK R., COHEN J., 1976. Further Considerations on the Methodological Analysis of Segregation Indices, *American Sociological Review,* Vol. 41, pp.630-37

COWGILL D., COWGILL M., 1951. An Index of Segregation Based on Block Statistics, *American Sociological Review,* Vol.16, pp. 825-831

DUNCAN D., DUNCAN B., 1955. A methodological analysis of segregation indexes, *American Sociological Review*, Vol. 20, No. 2, pp. 210-217.

FOSSETT M., 2017. Index Bias and Current Practices. In: *New Methods for Measuring and Analyzing Segregation. The Springer Series on Demographic Methods and Population Analysis*, Vol. 42, Cham:Springer.

GINI C., 1912. Sulla misura della concentrazione e della variabilità dei caratteri, *Atti del R. Istituto Veneto di Science, Lettere ed Arti*

JAHN J., 1950. The measurement of ecological segregation: derivation of an index based on the criterion of reproducibility, *American Sociological Review,* Vol. 15 pp.100-104

JAHN J., SCHMID C., SCHRAG C., 1947. The Measurement of Ecological Segregation, *American Sociological Review,* Vol. 12, pp.293-303

MASSEY D. S., DENTON N. A., 1988. The dimensions of residential segregation, *Social forces*, Vol. 67, No. 2, pp.281-315.

MAZZA A., PUNZO A., 2015. On the upward bias of the dissimilarity index and its corrections, *Sociological Methods and Research*, Vol. 44, pp. 80–107.

Ransom M. R. 2000. Sampling Distributions of Segregation Indexes, *Sociological Methods & Research*, Vol. 28 pp. 454-475.

TAEUBER, K. E., TAEUBER A. F., 1965. *Negroes in Cities: Residential Segregation and Neighborhood Change*. Chicago, IL: Aldine.

VOAS D., WILLIAMSON P., 2000. The Scale of Dissimilarity: Concepts, Measurement and an Application to Socio-Economic Variation across England

and Wales, *Transactions of the Institute of British Geographers*, Vol. 25, No. 4, pp. 465-481

WILLIAMS J., 1948. Another Commentary on So-called Segregation Indices, *American Sociological Review*, Vol. 13, pp. 298–303.

WINSHIP C., 1977. A re-evaluation of indices of residential segregation, *Social Forces*, Vol. 55, pp. 1058–1066.

WRIGHT J., 1937. Some measures of distribution, *Annals of the Association of American Geographers*, Vol. 27, pp. 177-211

SUMMARY

Dealing with the bias of the dissimilarity index of segregation

The dissimilarity index is widely used to evaluate the extent of segregation in the allocation of a minority group in two or more spatial units. There is a widespread awareness that due to its sensitivity to random allocation, it is inherently subject to an upward bias. This bias may be irrelevant when the sets concerned are large, but if the area population is small or the minority proportion is very low, the index can be highly misleading.

Common strategies used in literature to deal with the index bias rely on the use of informal rules of thumb, which at least have the side effect of restricting the scope of segregation studies.

In time, solutions for dealing directly with the index bias have been proposed. These solutions are based on adjusting the index scores downward, and require the computation of the sampling distribution of the index. Most of the methods proposed use computation-intensive techniques that have the drawback of introducing complexity and substantial computational burdens. The lack of user-friendly computer programs implementing these methods has strongly affected their wider adoption.

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