segregsmall: A command to estimate segregation in the presence of small units

Xavier D'Haultfœuille
CREST
Palaiseau, France
xavier.dhaultfoeuille@ensae.fr

Lucas Girard
CREST
Palaiseau, France
lucas.girard@ensae.fr

Roland Rathelot
University of Warwick
Coventry, UK
r.rathelot@warwick.ac.uk

Abstract. Suppose that a population, comprised of a minority and a majority group, is allocated into units, which can be neighborhoods, firms, classrooms, etc. Qualitatively, there is some segregation whenever the allocation process leads to the concentration of minority individuals in some units more than in others. Quantitative measures of segregation have struggled with the small-unit bias. When units contain few individuals, indices based on the minority shares in units are upward biased. For instance, they would point to a positive amount of segregation even when the allocation process is strictly random. The Stata command `segregsmall` implements three recent methods correcting for such bias: the nonparametric, partial identification approach of D'Haultfœuille and Rathelot (2017), the parametric model of Rathelot (2012), and the linear correction of Carrington and Troske (1997). The package also allows for conditional analyses, namely measures of segregation taking into account characteristics of the individuals or the units.

Keywords: segregation, small-unit bias, partial identification

1 Introduction

We consider a population made of two groups (minority and majority) whose individuals are spread across units. Units can be geographical areas, residential neighborhoods, firms, classrooms, or other clusters provided that every individual belongs to exactly one unit. We seek to measure the extent to which individuals from the minority group are concentrated in some units more than in others. Throughout the paper, we follow the literature and use the word "segregation" as a neutral term to refer to such concentration. Measuring the magnitude of segregation is a necessary step to understand the underlying mechanisms and design adequate policies.

A natural way to measure segregation is to start from the minority shares X_i/K_i, where X_i is the number of individuals from the minority group and K_i the number of individuals (or unit's size) in unit i, and then compute an inequality index based on the distribution of the proportions (X_i/K_i) across units.

There are two possible benchmarks to assess the magnitude of these indices. Evenness relates to the case where all minority shares X_i/K_i are equal across units. Randomness relates to the case where the underlying allocation process assigns minority individuals at random across units. If p_i is the probability that an arbitrary individual in unit i belongs to the minority, randomness means that probabilities p_i are equal across units i. Past research has stressed the difference between both benchmarks, especially when the units are of small size (Cortese et al. 1976). The minority share X_i/K_i is only an estimate of p_i, and even if $(p_i)_i$ are all equal, there will be some variation in $(X_i/K_i)_i$, especially if the units' sizes $(K_i)_i$ are small. If one is interested in the deviations from the randomness case, indices based on minority shares, which measure the deviation from evenness, will overestimate the level of segregation. This issue is known as small-unit bias.

The problem is pervasive in applied research. For workplace and school segregation, a large share of firms have less than ten employees and classrooms have usually between twenty and forty students. The bias also arises when the units are not small *per se* but only surveys of individuals are available. This is the case when one attempts to measure residential segregation using the local strata of households surveys.

Two main approaches have been proposed in the literature to deal with the small-unit bias. One strand proposes to correct the so-called naive inequality indices based on the minority shares $(X_i/K_i)_i$. The idea was initially proposed by Cortese et al. (1976) and Winship (1977) for the Duncan index. Carrington and Troske (1997, CT hereafter) extend the correction to other indices. Åslund and Skans (2009) adapt it to measure segregation conditional on covariates. Allen et al. (2015) develop another adjustment based on bootstrap. These corrections all aim at switching the benchmark from evenness to randomness, by subtracting an estimate of the bias from the initial, naive index.

Another approach, adopted by Rathelot (2012, R hereafter) and D'Haultfœuille and Rathelot (2017, HR hereafter), defines segregation using an inequality index based on the unobserved probabilities $(p_i)_i$, as a functional of the distribution F_p of p_i. In line with the rest of the literature, they assume that X_i are independently distributed in a $\text{Bin}(K_i, p_i)$. Conditional on K_i and p_i, R assumes a mixture of Beta distributions for F_p and derives the segregation index as a function of the parameters of the distribution. HR follow a nonparametric method leaving F_p unspecified; they show that the first moments of F_p are identified under the previous binomial assumption and obtain partial identification results on the segregation measure. Both R and HR construct confidence intervals for the segregation indices. HR also extend the methodology to study conditional segregation indices, namely measures of "net" or "residual" segregation taking into account other covariates (either of units or individuals) that may influence the allocation process.

The Stata command `segregsmall` allows social researchers to measure segregation in the context of small units. The command implements the methods proposed by R, HR, and CT. Conditional indices are available for all three methods. With R and HR, the command computes confidence intervals obtained by bootstrap. Finally, the command also implements a test of the binomial assumption.

This paper describes the command and presents the three methods it implements. Section 2 defines the set-up, the parameters of interest, and synthesizes the estimation and inference methods of R, HR, and CT. Section 3 details the syntax, options, stored results of the `segregsmall` command, and discusses its execution time. Section 4 presents an application of the command on French firm data to measure workplace segregation between foreigners and natives across workplaces. Section 5 concludes.

2 Set-up, estimation, and inference

2.1 The setting and the parameters of interest

The population studied is assumed to be split into two groups: a group of interest, henceforth the minority group, and the rest of the population. Individuals are distributed across units. For each unit, we assume that there exists a random variable p that represents the probability for any individual belonging to this unit to be a member of the minority. The total number of individuals in a unit is denoted by K.

We now introduce the segregation indices we focus on hereafter. We consider first unconditional indices; conditional indices are introduced in Section 2.6. Let us first assume that K is fixed. A segregation index θ is then a functional of the cumulative distribution function (cdf) F_p of p and of $m_{01} = \mathrm{E}(p)$, that is $\theta = g(F_p, m_{01})$.[1] Roughly speaking, one expects such an index to be minimal when F_p is degenerate, and maximal when $p \in \{0, 1\}$. In the former case, the probability of belonging to the minority is the same in all units, whereas in the latter case, the minority group is concentrated in a subset of units only.

The command `segregsmall` estimates five classical segregation indices satisfying this property, namely:

$$
\begin{aligned}
\mathrm{D} &= \frac{\int |u - m_{01}| \, \mathrm{d}F_p(u)}{2m_{01}(1-m_{01})} && \text{(Duncan)},\\
\mathrm{T} &= 1 - \frac{\int \{u\ln(u) + (1-u)\ln(1-u)\}\, \mathrm{d}F_p(u)}{m_{01}\ln(m_{01}) + (1-m_{01})\ln(1-m_{01})} && \text{(Theil)},\\
\mathrm{A(b)} &= 1 - \frac{m_{01}^{\frac{-b}{1-b}}}{1-m_{01}} \left\{\int (1-u)^{1-b} u^b \, \mathrm{d}F_p(u)\right\}^{\frac{1}{1-b}} && \text{(Atkinson with } b \in (0,1)),\\
\mathrm{CW} &= \frac{\int (u - m_{01})^2 \, \mathrm{d}F_p(u)}{m_{01}(1-m_{01})} && \text{(Coworker)},\\
\mathrm{G} &= \frac{1 - m_{01} - \int F_p^2(u) \ \mathrm{d}u}{m_{01}(1-m_{01})} && \text{(Gini)}.
\end{aligned}
$$

When K is random and takes values in $\mathcal{K}$, θ is defined as a weighted average of indices conditional on $K = k$, denoted $\theta^k = g(F_p^k, m_{01}^k)$ with F_p^k the cdf of p conditional on $K = k$, and $m_{01}^k = \mathrm{E}[p|K = k]$. Whether we study segregation at the unit-level or

[1] Such a notation may seem redundant since m_{01} already depends on F_p, but the reason why we make the dependence on m_{01} explicit will become clearer below.

at the individual-level matters for the weights used. The unit-level index θ_u satisfies

$$\theta_u = \sum_{k \in \mathcal{K}} \Pr(K = k)\theta^k, \tag{1}$$

whereas the individual-level segregation index θ_i is defined by

$$\theta_i = \sum_{k \in \mathcal{K}} \frac{k \Pr(K = k)}{\mathrm{E}(K)} \theta^k. \tag{2}$$

To estimate θ, we assume hereafter that the researcher has at her disposal K; however the probability p remains unobserved. Instead, she only observes X, the number of individuals belonging to the minority in the unit. By definition of p, we have $\mathrm{E}[X|K, p] = Kp$, which implies that the proportion of individuals from the minority, X/K, is an unbiased estimator of p. However, because it varies conditional on p, X/K is more dispersed than p. As a result, we have for usual segregation indices including the five ones above,

$$g\left(F_{X/K}, m_{01}\right) > g(F_p, m_{01}) = \theta.$$

In other words, even in the absence of statistical uncertainty on the distribution of X/K, we would still overestimate the segregation index by using X/K in place of p. Moreover, this bias increases as K decreases. We refer to this issue as the small-unit bias hereafter.

The binomial assumption We assume henceforth that individuals are allocated into units independently from each other. Namely, X is assumed to follow, conditional on p and K, a binomial distribution $\mathrm{Bin}(K, p)$. This hypothesis may be restrictive when the allocation process is in some way sequential and influenced by the composition of units. Importantly, this assumption is testable (see Section 2.5).

2.2 Nonparametric approach

Identification This approach, followed by HR, leaves the distribution F_p of p unrestricted. Combined with the binomial assumption, it entails a nonparametric binomial mixture model for X. Let us first suppose that K is constant; if not, we can simply retrieve aggregated indices θ_u and θ_i using (1) and (2). We also assume that $K > 1$; if $K = 1$, the distribution of X is not informative on θ and we only get trivial bounds on it, namely 0 and 1 for the five indices above.

First, some algebra yields a one-to-one mapping between the distribution of X, defined by the K probabilities $P_0 = (P_{01}, \ldots, P_{0K})'$ with $P_{0j} = \Pr(X = j)$, and the first K moments of F_p, denoted $m_0 = (m_{01}, \ldots, m_{0k})'$:

$$P_0 = Qm_0,$$

with Q the $K \times K$ matrix with generic entry (i, j) equal to $\binom{K}{j}\binom{j}{i}(-1)^{j-i}$.

It follows that m_0 is identified from the distribution of X, hence any parameter depending only on m_0 is point identified. It is the case of θ^{CW} as soon as $K \geq 2$. Second, there may be a single distribution F^* corresponding to m_0. This happens if (and only if) m_0 belongs to the boundary $\partial\mathcal{M}$ of the moment space $\mathcal{M}$.[2] Then F^* is a discrete distribution with at most $L+1$ support points, where L is the integer part of $(K+1)/2$. For instance, when $K=2$, $\mathcal{M} = \{(m_{01}, m_{02}) \in [0,1]^2 : m_{01}^2 \leq m_{02} \leq m_{01}\}$, since $\mathrm{V}(p) \geq 0$ and $p^2 \leq p$. Then $\partial\mathcal{M}$ corresponds to Dirac and Bernoulli distributions, for which we have respectively $\mathrm{V}(p) = 0$ and $p^2 = p$.

When m_0 belongs to the interior $\overset{\circ}{\mathcal{M}}$ of the moment space, there are infinitely many distributions F_p corresponding to m_0. Then, unless we consider θ^{CW}, θ is not identified in general. Nevertheless, HR show that the sharp identified set on θ can be computed in a relatively easy way under the following restriction.

Assumption 1. *$g(F, m_{01}) = \nu\left(\int h(x, m_{01})\,\mathrm{d}F(x), m_{01}\right)$, where h and ν are continuous and $\nu(\cdot, m_{01})$ is monotonic.*

Assumption 1 fails for the Gini but is satisfied by the Duncan, the Theil, the Atkinson, and the Coworker indices. Under this condition, the bounds on $\int h(x, m_{01})\,\mathrm{d}F(x)$, and thus on θ, are attained on distributions with no more than $K+1$ support points. Specifically, let $\mathcal{D}_{m_0}^{K+1}$ denote the set of distributions on $[0,1]$ with at most $K+1$ support points for which the vector of first K moments equals m_0. Then the sharp identified set on θ is $[\underline{\theta}, \overline{\theta}]$, with

$$\underline{\theta} = \inf_{F \in \mathcal{D}_{m_0}^{K+1}} g\left(F, m_{01}\right), \quad \overline{\theta} = \sup_{F \in \mathcal{D}_{m_0}^{K+1}} g\left(F, m_{01}\right). \tag{3}$$

The following theorem, which reproduces Theorem 2.1 of HR, summarizes the previous discussion. Hereafter, we let $\underline{\theta}$ and $\overline{\theta}$ denote the sharp lower and upper bounds on θ, whether or not θ is point identified.

Theorem 1. *– If $m_0 \in \partial\mathcal{M}$, $\underline{\theta} = \overline{\theta} = g(F^*, m_{01})$, where F^* is the unique cdf for which the first K moments are equal to m_0. Moreover, F^* has at most $L+1$ support points.*
– If $m_0 \in \overset{\circ}{\mathcal{M}}$ and Assumption 1 holds, $\underline{\theta}$ and $\overline{\theta}$ are defined by (3).

In the interior case, computing the bounds still requires a nonlinear optimization under constraints that are also nonlinear in the support points. Yet, the problem can be further simplified under additional assumption using the theory of Chebyshev systems. In particular, it requires that the function h in Assumption 1 does not depend on m_{01}, a condition satisfied by the Theil and Atkinson indices. Basically, for those two indices, no numerical optimization is needed to compute the bounds $\underline{\theta}$ and $\overline{\theta}$. The idea behind is that the bounds are attained by two special discrete distributions, called *principal representations*. The interest is that finding the principal representations boils down to obtaining the roots of specific polynomials, which is much simpler and faster than solving (3). We refer to HR for more details on that matter.

[2] This claim and several others of this section are proved in Krein and Nudelman (1977).

Estimation Let us assume to have in hand an i.i.d. sample $(X_i)_{i=1,\dots,n}$ of n units, with constant sizes equal to $K > 1$. Theorem 1 shows that θ is either point or partially identified, depending on whether $m_0 \in \partial\mathcal{M}$ or $m_0 \in \mathring{\mathcal{M}}$. We follow this result to estimate $(\underline{\theta}, \overline{\theta})$. In a first step, we estimate P_0, and thus $m_0 = Q^{-1}P_0$, by constrained maximum likelihood. The constraints come from the binomial mixture model: $P_0 \in \mathcal{P} = \{Qm : m \in \mathcal{M}\}$. To compute the constrained MLE, HR show Lemma 1 below. Let us define $N_k = \sum_{i=1}^{n} \mathbb{1}\{X_i = k\}$, $\mathcal{S}_{L+1} = \{(x_1, \dots, x_{L+1}) : 0 \leq x_1 < \dots < x_{L+1} \leq 1\}$ and $\mathcal{T}_{L+1} = \{(y_1, \dots, y_{L+1}) \in [0,1]^{L+1} : \sum_{k=1}^{L+1} y_k = 1\}$.

Lemma 1. *The constrained MLE $\widehat{P} = (\widehat{P}_1, \dots, \widehat{P}_K)'$ satisfies*

$$\widehat{P}_k = \binom{K}{k} \sum_{j=1}^{L+1} \widehat{y}_j \widehat{x}_j^k (1 - \widehat{x}_j)^{K-k}, \quad \forall\, k \in \{1, \dots, K\},$$

where $\widehat{x} = (\widehat{x}_1, \dots, \widehat{x}_{L+1})$ and $\widehat{y} = (\widehat{y}_1, \dots, \widehat{y}_{L+1})$ are given by

$$(\widehat{x}, \widehat{y}) = \underset{(x,y) \in \mathcal{S}_{L+1} \times \mathcal{T}_{L+1}}{\text{argmax}} \sum_{k=0}^{K} N_k \ln \left\{ \sum_{j=1}^{L+1} y_j x_j^k (1 - x_j)^{K-k} \right\}.$$

In a second step, we estimate $(\underline{\theta}, \overline{\theta})$. First, we check whether $\widehat{m} \in \partial\mathcal{M}$. A simple possibility to do so is testing whether the unconstrained MLE $\widetilde{P} = (\widetilde{P}_1, \dots, \widetilde{P}_K)'$ satisfies $\widetilde{P} = \widehat{P}$ (in which case $\widehat{m} \in \mathring{\mathcal{M}}$ with probability approaching one) or not. Note that the unconstrained MLE simply satisfies $\widetilde{P}_k = N_k/n$ for all k.

When $\widetilde{P} \neq \widehat{P}$, we simply let $\widehat{\underline{\theta}} = \widehat{\overline{\theta}} = g(\widehat{F}, \widehat{m}_1)$, where $\widehat{F}$ is the distribution corresponding to $(\widehat{x}, \widehat{y})$. We refer to this situation as the *constrained case*. If $\widetilde{P} = \widehat{P}$, there are infinitely many distributions corresponding to $\widehat{m}$ and we estimate bounds for θ. We refer to this situation as the *unconstrained case*. For the Theil and Atkinson indices, the estimated bounds are obtained from the principal representations computed from $\widehat{m}$. For the Duncan index, optimization is required to obtain the estimated bounds. We obtain estimators of $\underline{\theta}$ and $\overline{\theta}$ by solving the optimization problems (3), replacing m_0 by its estimator $\widehat{m}$. Finally, the Coworker index only depends on (m_{01}, m_{02}). Thus, whether or not $\widetilde{P} = \widehat{P}$, this index can be estimated directly by replacing (m_{01}, m_{02}) by $(\widehat{m}_1, \widehat{m}_2)$.

Inference When Assumption 1 holds, HR show that the estimators of the bounds are consistent: $(\widehat{\underline{\theta}}, \widehat{\overline{\theta}}) \xrightarrow{P} (\underline{\theta}, \overline{\theta})$ as the number of units n tends to infinity. Under additional assumptions, HR characterize their asymptotic distributions. This enables to build valid asymptotic confidence intervals (CIs) for the index θ using a modified bootstrap procedure. The construction needs to take into account the fact that the lower bound and upper bound collapse when $m_0 \in \partial\mathcal{M}$ (point-identification) whereas they differ when $m_0 \in \mathring{\mathcal{M}}$ (partial identification). The underlying idea relates to the construction of CIs in the case of partial identification (see Imbens and Manski (2004), Stoye (2009)). HR define a confidence interval for the interior case, where only one of the two ends of the interval matters in the asymptotic coverage, and another for the boundary case. In order to obtain the nominal asymptotic coverage in all situations, HR define the final confidence interval by selecting one of them according to the length of the estimated identification interval $(\widehat{\overline{\theta}} - \widehat{\underline{\theta}})$ relative to sampling error.[3]

Random unit size The previous identification and estimation results can be adapted to cases where K is random and takes values in $\mathcal{K}$. Using the definitions of θ_u and θ_i in (1) and (2), the idea is to reason conditional on the unit size to get each θ^k, $k \in \mathcal{K}$, and replace the theoretical weights by plug-in estimators. More precisely, let $\widehat{\underline{\theta}}^k$ and $\widehat{\overline{\theta}}^k$ denote the estimators of the bounds of θ^k based on the subsample of units of size k. Let $\widehat{\Pr}(K = k) = n^{-1} \sum_{i=1}^n \mathbb{1}\{K_i = k\}$ and $\widehat{\mathrm{E}(K)} = n^{-1} \sum_{i=1}^n K_i$. Then the estimators of the bounds on θ_u and θ_i satisfy

$$\widehat{\underline{\theta}}_u = \sum_{k \in \mathcal{K}} \widehat{\Pr}(K = k) \widehat{\underline{\theta}}^k, \quad \widehat{\overline{\theta}}_u = \sum_{k \in \mathcal{K}} \widehat{\Pr}(K = k) \widehat{\overline{\theta}}^k,$$

$$\widehat{\underline{\theta}}_i = \sum_{k \in \mathcal{K}} \frac{k\widehat{\Pr}(K = k)}{\widehat{\mathrm{E}(K)}} \widehat{\underline{\theta}}^k, \quad \widehat{\overline{\theta}}_i = \sum_{k \in \mathcal{K}} \frac{k\widehat{\Pr}(K = k)}{\widehat{\mathrm{E}(K)}} \widehat{\overline{\theta}}^k.$$

Remark that as soon as for one size k the index θ^k is not point identified, the resulting aggregated index will be partially identified too. In other words, point identification of θ_u or θ_i requires to be in the constrained case for each $k \in \mathcal{K}$. This is unlikely to happen when the support of K contains very small sizes k, typically lower than 10.

Similar to the constant unit case, confidence intervals for the aggregated indices θ_u and θ_i are constructed by the modified bootstrap procedure detailed in HR. The randomness of K just involves an additional step that consists in drawing K in its empirical distribution.

Assuming independence between K and p The previous estimation and inference procedures are fully agnostic as regards possible dependence between K and p, which is a safe option when unit size may be a potential determinant of segregation. However, if one is ready to impose independence between these two variables, the identified bounds on $\theta_u = \theta_i$ get closer to each other. This is because the F_p^k coincide with the unconditional distribution of p. Thus, we can gather all units and identify the first $\overline{K}$ moments of F_p, with $\overline{K} = \max(\mathcal{K})$. Estimation and inference are performed as in the case of constant unit size, with K replaced by $\overline{K}$. Thus, assuming independence between K and p leads to an improvement in identification, since we identify more moments for most of the data than. It also leads to more accurate estimators, since one estimates a single

[3] Essentially, when this length is large (resp. small) relative to sampling error, the uncertainty related to partial identification (resp. to sampling) prevails and the interior-type (resp. boundary-type) confidence interval is used.

vector P on the whole sample, instead of doing so on each subsample $\{i : K_i = k\}$, for all $k \in \mathcal{K}$. An important particular case occurs when only some individuals in the unit are observed (e.g., survey data). Imagine units are of size $(K_i)_{i=1,\dots,n}$ but that, for each unit i, only $n_{K,i}$ individuals are sampled and observed. We let X_i denotes the number of individuals belonging to the reference group in this subgroup of $n_{K,i}$ people. As previously, X_i follows a binomial distribution $\text{Bin}(n_{K,i}, p_i)$ conditional on p_i and $n_{K,i}$. The previous results apply by simply replacing the unit size K by the number n_K of individuals observed in each unit. Moreover, in such settings, it is usually plausible to assume that the random variable n_K is independent of p conditional on the unit size as n_K depends on the survey process which, *a priori*, is orthogonal to the segregation phenomenon.

2.3 Parametric approach

This approach, followed by R, is similar to that of HR, except that it imposes a parametric restriction on F_p. Specifically, it is supposed to be a mixture of Beta distributions. Combined with the binomial assumption for the conditional distribution of X, the model becomes fully parametric and thus can be estimated by maximum likelihood. The indices are therefore point identified, contrary to the nonparametric approach of HR.

A concern might be that the parametric restriction leads to invalid results when the model is misspecified. However, R shows through simulations that segregation indices associated with various distributions, both continuous and discrete, are accurately proxied by his parametric approach.

Estimation and inference As in HR, we first assume that K is constant. Let $B(\cdot,\cdot)$ denote the beta function, c the number of components of the beta mixture, $v = (\alpha_j, \beta_j, \lambda_j)_{j=1,\dots,c}$ the vector of parameters with $(\alpha_j, \beta_j) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$ the two shape parameters of the j-th Beta distribution and $\lambda_j \in [0,1]$ its weight ($\sum_{j=1}^{c} \lambda_j = 1$). The probability density function of p distributed as a c-component mixture of Beta distributions with parameters v is:

$$f_v(t) = \sum_{j=1}^{c} \lambda_j \frac{t^{\alpha_j - 1}(1-t)^{\beta_j - 1}}{B(\alpha_j, \beta_j)}, \quad \forall t \in [0,1].$$

In this model, the probability that k individuals belong to the minority group can be written, after some algebra, as:

$$\Pr_v(X = k) = \binom{K}{k} \sum_{j=1}^{c} \lambda_j \frac{B(\alpha_j + k, \beta_j + K - k)}{B(\alpha_j, \beta_j)}.$$

Thus, the log-likelihood satisfies, up to terms independent of the parameter v,

$$\ell(v) = \sum_{k=0}^{K} N_k \times \ln \left\{ \sum_{j=1}^{c} \lambda_j \frac{B(\alpha_j + k, \beta_j + K - k)}{B(\alpha_j, \beta_j)} \right\},$$

Maximizing $v \mapsto \ell(v)$ yields the maximum likelihood estimator $\widehat{v}$. Using the parametric assumption on F_p, $\widehat{v}$ translates into an estimator $\widehat{F}_p$ of the distribution of p, which in turn yields an estimator $\widehat{\theta}$ of θ. The explicit expressions of the five indices above, as functions of the parameter v, are given in Appendix 7.1. Inference can be achieved by the delta method or by the bootstrap, performed at the unit level.

Random unit size The adaptation to this case is exactly similar to HR method. For each $k \in \mathcal{K}$, the MLE of θ^k is obtained using the subsample of units of size k. The weights are estimated by their empirical counterparts. The estimated aggregated indices are then obtained by plug-in, using (1) and (2). When K and p are assumed independent, all units can be pooled, independently of their size, to compute the MLE of v for the whole sample. As above, the resulting estimator $\widehat{v}$ allows us to estimate the distribution of p, and then θ.

2.4 Correction of the naive index

The approaches of HR and R are immune to the small-unit bias as they directly estimate $g(F_p, m_{01})$. Other, previous approaches rather start from the naive index $\theta_N = g(F_{X/K}, m_{01})$ and attempt to modify it, so that the parameter becomes less sensitive to changes in K. We present here the correction proposed by CT, which is the most popular in applied work.

CT's correction relies on the distinction between the randomness and evenness benchmarks, introduced notably by Cortese et al. (1976) and Winship (1977). Evenness corresponds to X/K being constant, whereas randomness refers to the case where p is constant. Under the binomial model, however, evenness cannot occur. The central idea of CT is then to convert θ_N, which measures departure from evenness, into a distance to randomness. Let θ_N^{ra} denote $g(F_{X^{\text{ra}}/K}, m_{01})$, where $X^{\text{ra}}|K \sim \text{Bin}(K, E(X/K))$. X^{ra}/K is the proportion we would observe if p was constant and equal to $E(p) = E(X/K)$. Then, assuming that $\theta \in [0,1]$, a constraint satisfied by the five indices above, CT's correction θ_{CT} is defined by $\theta_{CT} = (\theta_N - \theta_N^{\text{ra}})/(1 - \theta_N^{\text{ra}})$. CT suggest the following simulation-based estimator of θ_{CT}. Let $\widehat{E}(p)$ denote the sample average of X/K. For $s = 1, ..., S$, draw $X_{i,s}^{\text{ra}} \sim \text{Bin}(K_i, \widehat{E}(p))$ independently for each unit i. Then, letting $\widehat{F}_s^{\text{ra}}$ and $\widehat{m}_{1,s}$ denote respectively the empirical distribution and mean of $(X_{i,s}^{\text{ra}}/K_i)_{i=1,...,n}$, compute $\widehat{\theta}_{N,s}^{\text{ra}} = g(\widehat{F}_s^{\text{ra}}, \widehat{m}_{1,s})$. The estimator of θ_N^{ra} is then the mean over the S replications, $\widehat{\theta}_N^{\text{ra}} = S^{-1}\sum_{s=1}^S \widehat{\theta}_{N,s}^{\text{ra}}$. Finally, $\widehat{\theta}_{\text{CT}} = (\widehat{\theta}_N - \widehat{\theta}_N^{\text{ra}})/(1 - \widehat{\theta}_N^{\text{ra}})$, with $\widehat{\theta}_N$ the plug-in estimator of θ_N. The quantiles of $(\widehat{\theta}_{N,s}^{\text{ra}})_{s=1,...,S}$ can be used to test that the data are consistent with random allocation using randomization tests (see Boisso et al. 1994, and CT).

Links with HR and R In general, $\theta_{CT} \neq \theta$. They do coincide however in the extreme cases of no segregation, where p is constant, and "full" segregation, where p follows a Bernoulli distribution. We refer to Section 2.3 of R and Section 2.4 of HR for further discussion on the relationship between θ_{CT} and θ.

2.5 Test of the binomial assumption

We have relied so far on the binomial assumption $X|K, p \sim \text{Bin}(K, p)$. This assumption implies that $P_0 \in \mathcal{P} = \{Qm : m \in \mathcal{M}\}$. A vector $(m_1, ..., m_K)$ in $\mathcal{M}$ has to satisfy some restrictions, such as $m_2 \geq m_1^2$ (i.e., non-negative variance). Hence, we could have $Q^{-1}P_0 \notin \mathcal{M}$ if the distribution of X conditional on K and p is not binomial. In other words, the binomial assumption is testable.

HR propose a likelihood ratio test of $P_0 \in \mathcal{P}$, where the constrained estimator under the null hypothesis is $\widehat{P}$, whereas the unconstrained MLE is $\widetilde{P}$. Note that these estimators are already computed to estimate $(\underline{\theta}, \overline{\theta})$. For a unit size equal to k, the test statistic satisfies

$$LR_k = 2\sum_{x=0}^{k} N_x \ln\left(\frac{\widetilde{P}_x}{\widehat{P}_x}\right) = 2\sum_{x=0}^{k} N_x \ln\left(\frac{N_x}{n\widehat{P}_x}\right),$$

where we let $N_x \ln[N_x/(n\widehat{P}_x)] = 0$ if $N_x = 0$.

With a random unit size, the test statistic is then $LR_n = \sum_{k\in\mathcal{K}} \widehat{\Pr}(K = k)LR_k$, where in LR_k, $N_x = \sum_{i=1}^n \mathbb{1}\{K_i = k, X_i = x\}$. The critical values of the test are obtained by approximating the distribution of LR under the null by bootstrap. The bootstrap is performed as follows. First, we draw n units of sizes K_i^* in the empirical distribution of K. Second, we draw X_i^* according to $\widehat{P}^{K_i^*}$, where $\widehat{P}^k$ is the constrained MLE of P_0^k, the distribution of X conditional on $K = k$. The bootstrapped test statistic LR^* is then computed in the sample $(K_i^*, X_i^*)_{i=1,...,n}$, which is drawn under the null hypothesis. For a level $1-\alpha \in (0,1)$, the critical region of the test is defined by:

$$CR = \{LR > c_{1-\alpha}(LR^*)\},$$

with $c_{1-\alpha}$ the quantile of order $1-\alpha$ of LR^*.

The results of HR imply that the test has an asymptotic level equal to α and is consistent. Remark however that it tests $P_0 \in \mathcal{P}$, which is an implication of the binomial assumption, rather than this assumption itself. This means that the binomial assumption may fail but still, $P_0 \in \mathcal{P}$: $X|K, p$ could fail to be binomial, yet the distribution of X given K could be rationalized by a binomial mixture.

2.6 Conditional segregation indices

Conditional indices aim at accounting for the fact that part of the segregation along the minority/majority dimension may be driven by sorting according to other dimensions. In this sense, they measure the net or residual level of segregation, when the contribution of covariates to segregation is removed (see Åslund and Skans 2009). To illustrate this point, let us consider workplace segregation between foreigners and natives. Foreigners may be hired more in some sectors of the economy on the basis of sector-specific skills. Imagine an extreme case where, within each sector, all firms hire foreigners with the same probability. As long as these probabilities differ from one sector to another, an unconditional segregation index would be positive. On the contrary, the conditional index defined in (4) below would indicate no segregation as it controls for the influence of the sector, a characteristic of units, in the allocation process. Similarly, foreigners may be hired with the same probability for all low-skilled jobs (resp. all high-skilled jobs), but the probabilities for these two types of job may differ. In this case again, failing to account for this characteristic would lead to a positive unconditional index, while the conditional index defined in (5) below would indicate no segregation.

The previous discussion underscores that covariates can be defined either at the unit level or at the level of an individual/position. We separate the two cases below, as they lead to different treatments.

Unit-level covariates Let $Z \in \{1, \dots, \overline{Z}\}$ denote a characteristic of a unit, which is assumed to be discrete. To take into account Z in the allocation process, we measure segregation conditional on Z. For each $z \in \{1, \dots, \overline{Z}\}$, let θ_{0z} denote the segregation index we consider conditional on $Z = z$. The subscript 0 indicates that we consider a generic index of interest, which could correspond to either θ or θ_{CT}. Whatever the index, the estimation of θ_{0z} is done exactly as in the unconditional case, focusing on the subsample $\{i : Z_i = z\}$.

The index θ_{0z} can be of interest by itself. We can also consider an aggregate conditional index defined as follows:[4]

$$\theta_{0,\mathrm{u}}^{\mathrm{cond}} = \sum_{z=1}^{\overline{Z}} \Pr(Z = z)\theta_{0z}. \tag{4}$$

The estimation of $\theta_{0,\mathrm{u}}^{\mathrm{cond}}$ is obtained by plug-in, with $n^{-1}\sum_{i=1}^{n} \mathbb{1}\{Z_i = z\}$ the empirical counterpart of $\Pr(Z = z)$. For HR and R methods, a similar bootstrap procedure as in the random size case provides asymptotic confidence intervals for $\theta_{0,\mathrm{u}}^{\mathrm{cond}}$.[5]

Individual- or position-level covariates Let $W \in \{1, \dots, \overline{W}\}$ denote a characteristic of an individual or of a position. To resume the example of workplace segregation, a characteristic attached to individuals can be education, whereas a characteristic linked to positions can refer to the type of occupation (e.g., high-skilled versus low-skilled). While these two forms of covariates may lead to different interpretations, they are similar as regards estimation and inference.

[4] Note that for θ_{CT}, the aggregate conditional indices defined by (4), and similarly for (5) below, slightly differ from the conditional index of Åslund and Skans (2009). Broadly speaking, (4) and (5) aggregate the corrected indices computed conditional on each type while Åslund and Skans (2009) do one unique correction in order to directly obtain their conditional corrected index. The former has the advantage to be more general and notably can be used as such in HR and R approaches.

[5] The initial step of the bootstrap procedure becomes drawing units in the joint empirical distribution of (K, Z).

For each unit and each type $w \in \{1, \dots, \overline{W}\}$, we suppose to observe X_w and K_w, which are respectively the number of individuals with characteristic $W = w$ (or in positions satisfying $W = w$) who belong to the minority group, and the overall number of individuals (or positions) of type $W = w$ in the unit. As above, we define θ_{0w} as the segregation index of interest conditional on $W = w$. With individual- or position-level covariates, the idea is to consider the subsample of individuals (or positions) such that $W = w$, instead of a subsample of units. Hence, θ_{0w} can be estimated exactly as in the unconditional case simply using (X_w, K_w) instead of (X, K).[6]

Again, θ_{0w} might be a relevant parameter of interest on his own. Researchers can also be interested in an aggregated conditional index:

$$\theta_{0,\mathrm{i}}^{\mathrm{cond}} = \sum_{w=1}^{\overline{W}} \Pr(W = w)\theta_{0w}. \tag{5}$$

The estimation of $\theta_{0,\mathrm{i}}^{\mathrm{cond}}$ is obtained by plug-in, with $(\sum_{i=1}^{n} K_{wi})/(\sum_{i=1}^{n} K_i)$ the empirical counterpart of $\Pr(W = w)$. For HR and R methods, as previously, a modified bootstrap procedure provides asymptotic confidence intervals for $\theta_{0,\mathrm{i}}^{\mathrm{cond}}$.[7]

3 The segregsmall command

The **segregsmall** command is compatible with Stata 14.2 and later versions.

3.1 Syntax

The syntax of **segregsmall** is as follows:

```
segregsmall varlist [if] [in] , method(string) format(string) [
   conditional(string) withsingle excludingsinglepertype independencekp
   level(#) repbootstrap(#) noci testbinomial repct(#) atkinson(#)]
```

3.2 Description and main options

The command **segregsmall** estimates the five classical segregation indices mentioned above (Duncan, Theil, Atkinson, Coworker, and Gini) using D'Haultfœuille and Rathelot (2017), Rathelot (2012), or Carrington and Troske (1997) method. It provides confidence intervals obtained by bootstrap in the approaches of HR and R and allows for conditional analysis for all three methods.

[6] Remark that, in the general random sizes case without assuming $K \perp\!\!\!\perp p$, it makes more sense to consider the index θ_i that uses individual-level weights (compared to unit-level ones) because the types are defined at this individual-/position-level.

[7] The initial step of the bootstrap procedure becomes drawing units in the empirical distribution of units, hence keeping fixed the composition of the units with respect to W.

`method` specifies the method used. Its argument must be one of: *np*, *beta*, *ct*. Argument *np*, standing for nonparametric, implements HR method. The command does not report the Gini index in this case as it does not verify Assumption 1. The choice *beta* implements R's method assuming a Beta distribution for F_p.[8] Both methods provide estimates of the same parameters of interest, namely θ if K is fixed and, unless `independencekp` is specified, (θ_u, θ_i) if K is random. By default, they report asymptotic confidence intervals obtained by bootstrap. With the argument *ct*, the command estimates the naive and CT-corrected indices θ_N and θ_{CT}. Confidence intervals are not computed for these parameters.

`format` indicates the format of the dataset used and needs to be either *unit* (datasets where an observation is a unit) or *indiv* (datasets where an observation is an individual). The option determines the variables to be put in *varlist*. For unconditional analyses (the default without option `conditional`), these are:

- `K X` for unit-level datasets, `K` and `X` correspond to the variables K and X introduced in Section 2: the number of individuals and the number of minority individuals. `K` has to be strictly positive integers and `X` positive or null integers. `X` should be lower or equal to `K` for each unit.

- `id_unit I_minority` for individual-level datasets, `id_unit` is the identifier of the unit the individual belongs to. `I_minority` is a dummy variable equal to 1 when the individual belongs to the minority group, 0 otherwise.

`conditional` this option triggers the computation of conditional segregation indices. Its arguments must be either *unit* or *indiv* and it specifies the level at which are defined the covariates included in the analysis. For conditional analysis, *varlist* has to be:

- `K X Z` for unit-level datasets, or `id_unit I_minority Z` for individual-level datasets, with covariates defined at unit-level (*unit*). The variables `K`, `X`, `id_unit`, and `I_minority` are the same as in unconditional analyses. `Z` corresponds to the variable Z, the characteristics of units defined in Section 2.6. `Z` needs to take values in $\{1, 2, \dots, \overline{Z}\}$ with $\overline{Z} \geq 2$.

- `id_unit I_minority W` for individual-level datasets with covariates defined at the level of individuals or any sub-unit level (*indiv*). `W` corresponds to the variable W, the individual (or position) characteristics introduced in Section 2.6. `W` has to take values in $\{1, 2, \dots, \overline{W}\}$ with $\overline{W} \geq 2$.

3.3 Additional options

`withsingle` includes single units (with only one individual) in the analysis. As explained in Section 2.2, single units are in general uninformative about the level of segregation.

[8] R assumes a mixture of Beta distributions. However, simulations reveal that the differences between the indices obtained with a two or higher component mixture versus a simple Beta are marginal in most cases, `segregsmall` uses a Beta assumption for simplicity. Also, the command allows to assess the reliability of this restriction since the indices obtained with the beta restriction can be compared with the nonparametric estimates that leaves F_p unrestricted.

By default, they are not included in the data used. The option is available both for unconditional or conditional analyses.

`excludingsinglepertype` excludes single cells (unit × type) from the analysis. The option is only relevant and available in conditional analyses with covariates defined at the individual/sub-unit level. In this setting, the role of a unit in unconditional analyses is played by a cell defined as the intersection of a unit and an individual type (see Section 2.6). As just described, *units* with only one individual are dropped by default. Yet, this does not prevent the existence of single *cells* coming from units with more than one individual but that have only one individual of a given type $W = w$. Without option `excludingsinglepertype`, those single cells are included in the analysis, which can lead to broad estimated identification sets in HR method, all the more so as the number $\overline{W}$ of types is large. With the option, they are dropped. For consistency, the options `withsingle` and `excludingsinglepertype` are mutually exclusive.

`independencekp` assumes independence between K and p. The option is only available with *np* and *beta* methods.

`level` sets the confidence level, which has to be a scalar in $(0, 1)$. With *np* and *beta* methods, by default, the traditional 90%, 95%, and 99% confidence levels are saved (see Section 3.4) and the 95% confidence interval is displayed in Stata output. The option permits to save and display a personalized level besides (the other three are still stored). With *ct* method, by default, the empirical quantiles of the index under random allocation are stored for the orders 0.01, 0.05, 0.10, 0.90, 0.95, and 0.99. The option additionally saves the empirical quantiles at order τ and $1 - \tau$ with τ the argument of the option.

`repbootstrap` specifies the number of bootstrap iterations used to construct confidence intervals in *np* and *beta* methods. The default number is 200. It is also the number of bootstrap repetitions used to test the binomial assumption.

`noci` restricts the command to estimation: confidence intervals are not computed. The option is only applicable to *np* and *beta* methods.

`testbinomial` implements the test of the binomial assumption. More precisely, with `method(`*np*`)` and without options `independencekp` nor `noci`, the test is made by default and saved: the option only displays the result in Stata output. In any other situations (*beta* or *ct* methods, no CIs, or assuming $K \perp\!\!\!\perp p$), the option performs the test in addition to estimation and potential inference. In both cases, the number of bootstrap repetitions used for the test is the same as the one specified by option `repbootstrap`. When the user wants to test the binomial assumption, we recommend always to do so combined with inference using HR method in the general case (namely, without assuming independence between K and p): together with the test, it will give estimation and confidence intervals from *np* method virtually for free. The option is only available in unconditional analyses.[9]

[9] The test of the assumption type by type can be done manually by restricting the sample used through the options [*if*] [*in*].

`repct` sets the number S of draws used to estimate θ_N^{ra} in CT's correction. Its argument needs to be a positive integer. The default value is 50.

`atkinson` allows the user to specify the parameter b of the Atkinson index. Its argument has to be a real in $(0, 1)$. The default value is 0.5; it is the only one that ensures the symmetry property for the Atkinson index (i.e., the index does not change when swapping the minority/majority labels).

3.4 Saved results

The objects saved by `segregsmall` depend on the options, in particular, whether the analysis is unconditional or conditional. They can be gathered into three types of information about: *(i)* the data included in the analysis, *(ii)* the method and assumptions used, *(iii)* the estimation and inference results.

In this section, we list the objects saved in `e()` by the command and detail their contents when they relate to estimation and inference results. The remaining objects have self-explanatory names and are described in the help page of the `segregsmall` command.

Data included in the analysis Below, names with prefix `I_` denote dummy variables equal to 1 if what follows is true, 0 otherwise. Objects stored in unconditional analyses are printed in black. *Additional* objects stored in conditional analyses are displayed in gray. The superscript *u indicates that the objects are only relevant and saved for unit-level covariates, the superscript *i for the individual-level covariates.

Scalars:

`e(I_withsingle)`

`e(I_excludingsinglepertype)`

`e(I_unit_level_characteristic)`

`e(nb_types)`

`e(nb_units_total)`

`e(nb_units_single)`

`e(nb_units_studied)`

`e(nb_individuals)`

`e(nb_minority_individuals)`

`e(prop_minority_hat)`

`e(K_max)`

`e(nb_K_with_obs)`

`e(nb_cells_studied_sum_across_type)`^{*i}

`e(nb_single_cells_sum_across_type)`^{*i}

Matrices:

`e(list_K_with_obs)`

`e(type_frequencies)`

`e(type_probabilities)`

`e(summary_info_data_per_type)`

`e(nb_units_studied_per_type)`^{*u}

`e(nb_cells_studied_per_type)`^{*i}

Method used

Scalars:

`e(I_method_np)`

`e(I_method_beta)`

`e(I_method_ct)`

`e(I_conditional)`

`e(I_unit_level_characteristic)`

`e(I_hyp_independenceKp)`

`e(I_noci)`

`e(nb_bootstrap_repetition)`

`e(specified_level)`

`e(I_testbinomial)`

`e(nb_ct_repetition)`

`e(b_atkinson)`

Estimation and inference Objects relative to unconditional analyses are in black (left-hand column); those relative to conditional analyses are in gray (right-hand column). Superscripts $*np$ and $*beta$ indicate that objects are only relevant and saved with *np* and *beta* method.

Scalars:

`e(I_constrained_case)`^{*np}

`e(I_constrained_case)`^{*np}

Matrices:

`e(estimates_ci)`

`e(info_distribution_of_p)`^{*np,*beta}

`e(test_binomial_results)`

`e(I_constrained_case_per_type)`^{*np}

`e(estimates_ci_aggregated)`

`e(estimates_ci_type_#)`

The matrices whose name includes `estimates_ci` store the results of estimation and possible inference. The content of `e(estimates_ci)` varies with the method used but its structure remains similar. Each row corresponds to an index.

With *beta* method, ten rows represent the two possible aggregated indices θ_u (unit-level weights) and θ_i (individual-level weights), when K is considered as random, for each of the five indices (Duncan, Theil, Atkinson, Coworker, and Gini). For each possible index weights × mapping, the columns store the estimated index using R method with a Beta distribution restriction on F_p, and asymptotic confidence intervals at the traditional 90%, 95%, and 99% levels (plus the one specified by `level` if any).

With *np* method, the rows are identical but there are only eight parameters since the Gini indices are absent. For each possible index weights × mapping, the columns of `e(estimates_ci)` save: the estimated bounds $\widehat{\underline{\theta}}_u$ and $\widehat{\overline{\theta}}_u$ for unit-level weights (or $\widehat{\underline{\theta}}_i$ and $\widehat{\overline{\theta}}_i$ for individual-level weights); a dummy variable equal to 1 if the confidence interval used is the boundary-case interval and 0 for the interior-case; the resulting asymptotic CI at the classical 90%, 95%, and 99% levels (plus the one specified by `level` if any).[10]

In conditional analyses, either with unit- or individual/position-level covariates, the matrices `e(estimates_ci_aggregated)` and `e(estimates_ci_type_#)` store exactly the same information as `e(estimates_ci)`: the former for the aggregated conditional index $\theta_{0,\mathrm{u}}^{\mathrm{cond}}$ or $\theta_{0,\mathrm{i}}^{\mathrm{cond}}$, the latter for the index conditional on a given type #, that is θ_{0z} with

[10] The boundary/interior CIs were discussed briefly in Section 2.2 §Inference. We refer to the original paper for further details.

unit-level characteristics or θ_{0w} with individual/position-level characteristics (# ranges from $z = 1$ to $\overline{Z}$ or $w = 1$ to $\overline{W}$).

With *ct* method, five rows correspond respectively to the Duncan, the Theil, the Atkinson, the Coworker, and the Gini indices. In columns: the naive index θ_N; the index under random allocation $\widehat{\theta}_N^{\text{ra}}$; the CT-corrected index θ_{CT}; the empirical standard deviation of the draws $(\widehat{\theta}_{N,s}^{\text{ra}})_{s=1,...,S}$ under random allocation; the "standardized score" originally proposed by Cortese et al. (1976), namely $(\theta_N - \widehat{\theta}_N^{\text{ra}})$ divided by that standard deviation; the empirical quantiles of $(\widehat{\theta}_{N,s}^{\text{ra}})_s$ at the orders: 0.01, 0.05, 0.10, 0.90, 0.95, 0.99 (τ and $1 - \tau$, with τ the argument of option `level` if this option is used).

`e(I_constrained_case)` is a dummy equal to 1 in the constrained case, 0 otherwise. As discussed in Section 2.2, with random unit size, it requires to be in the constrained case ($\mathcal{D}_{\widehat{m}}$ restricted to a singleton) for each size $k \in \mathcal{K}$. In this case, *np* method yields point-estimates for all indices. `e(I_constrained_case)` is identical in conditional analyses. The dummy is equal to 1 provided we are in the constrained case for each type. Otherwise, $\theta_{0,\text{u}}^{\text{cond}}$ and $\theta_{0,\text{i}}^{\text{cond}}$ are only partially-identified with *np* method.

`e(test_binomial_results)` is stored when the test of the binomial assumption is performed (see option `testbinomial`). It is a row vector whose first element saves the value of the test statistic LR_n and the second the p-value of the test where the null hypothesis is the binomial assumption.

np and *beta* methods save `e(info_distribution_of_p)` in unconditional analyses.[11] This matrix contains the information learned about the distribution of p in the estimation. In the general case, without assuming $K \perp\!\!\!\perp p$, it means the information as regards the conditional distributions F_p^k, for each $k \in \mathcal{K}$. With option `independencekp`, it is about the unconditional distribution F_p.

With *beta* option, all the $(F_p^k)_{k\in\mathcal{K}}$ (or F_p when assuming $K \perp\!\!\!\perp p$) are supposed to follow a Beta distribution. In the general case, `e(info_distribution_of_p)` is a matrix with $|\mathcal{K}|$ rows. Each row is associated with a size k and the columns report: the size k; the number of units of size k in the data used i.e., $\sum_{i=1}^n \mathbb{1}\{K_i = k\}$; the latter quantity expressed as a proportion over the n units studied; the number of components of the Beta mixture considered (that is 1); and the maximum likelihood estimators $\widehat{\alpha}_1$ and $\widehat{\beta}_1$ of the two shape parameters characterizing the Beta distribution assumed for F_p^k. In the case where $K \perp\!\!\!\perp p$ is supposed, the matrix `e(info_distribution_of_p)` is similar but consists of a single row as only one estimation is done pooling all units together. It contains the maximal size $\overline{K}$, the number of units n used for the estimation, and the estimates of the parameters that characterize the Beta distribution assumed for F_p.

With *np* option, the structure of `e(info_distribution_of_p)` is more involved for the approach is nonparametric. Without the restriction $K \perp\!\!\!\perp p$, it contains $3 \times \overline{K}$ rows and should be read by blocks of three rows. The k-th block concerns F_p^k. The first line shows some general information, namely the size k, the number of units of size k,

[11] In conditional analyses, the information can be retrieved manually for each type by restricting the sample used.

and the proportion of such units within the data used (as in *beta* method). The most important element is displayed in the fourth column and consists of a dummy variable equal to 1 if we are in the constrained case for F_p^k, that is $\widehat{m} \in \partial\mathcal{M}$ conditional on $K = k$. In this case, despite the nonparametric approach, the constrained maximum likelihood estimation yields an estimate $\widehat{F}$ of F_p^k which turns out to be a discrete distribution with at most $\lfloor (k+1)/2 \rfloor + 1$ support points (see Section 2.2 §Estimation). In this situation, the fifth column of the first row, within the three-row block, indicates the number of support points of $\widehat{F}$ and the two following rows characterize $\widehat{F}$ by reporting its support points and the corresponding probabilities. In the unconstrained case, the dummy is 0 and the two last rows, within the three-row block, are empty as there is no estimate of F_p^k then. When assuming $K \perp\!\!\!\perp p$, the matrix `e(info_distribution_of_p)` is analogous but is made of a single three-row block as it only deals with the unconditional distribution F_p. In this case (see Section 2.2 §Assuming independence between K and p), the estimation uses the first $\overline{K}$ moments of F_p. It is likely to fall in the constrained case since $\overline{K}$ will exceed 10 in most applications, a size above which simulations reveal that the probability to be in the constrained case is close to one even with large sample sizes n.

`e(info_distribution_of_p)` is interesting because virtually any segregation index is a functional of the distribution F_p (of the conditional distributions $(F_p^k)_k$ in general when taking into account the randomness of K). Consequently, an estimate of F_p (respectively of the $(F_p^k)_k$) enables to recover any other personalized segregation index.

3.5 Execution time

The times reported below are average over 50 repetitions on a desktop computer run under Windows 10 Enterprise with an Intel(R) Core(TM) i5-6600 CPU 3.30GHz processor (RAM 16 Go). The operations of `segregsmall` can be decomposed into a preparation stage and a stage devoted to estimation and inference.

Preparation stage The preparation stage is common to the three methods and reshapes the dataset. Its execution time is quick compared to the whole command and increases in the number n of units. For instance, with unit-level datasets, for K taking values in $\mathcal{K} = [5, 15]$, it lasts around 0.06 second with $n =$ 1,000, and 0.99 second with $n =$ 300,000. In conditional analyses, the execution time is approximately multiplied by the number of types: for example, 6.03 seconds for 5 types and 9.17 seconds for 10 types, with $\mathcal{K} = [5, 15]$ and $n =$ 300,000. With individual-level datasets, the preparation stage is longer since it is necessary first to form the units. With $\mathcal{K} = [5, 15]$, it takes 0.24 seconds with 1,000 units and 9.99 seconds with 300,000 units.

Estimation and inference stage The subsequent operations depend on the method used. The central brick of *np* and *beta* methods is the estimation of the indices for a given dataset (original or bootstrapped). The construction of CIs repeats the operation for each bootstrapped dataset. The execution time is thus more or less linear in the number of bootstrap repetitions (fixed by option `repbootstrap`). *ct* method requires to reshuffle the data under the randomness benchmark, hence an execution time broadly linear in the number of draws (controlled by option `repct`). Table 1 illustrates this dependence for *np* and *beta* methods as well as the effect of option `conditional`. Regarding the latter, for all three methods, the same operations as in unconditional analyses are done for each type (see Section 2.6). As a consequence, the execution time of `segregsmall` is roughly linear in the number of types included in the analysis.

Table 1: Execution time in seconds. Setting: unit-level datasets, $n = 300{,}000$, $\mathcal{K} = [5, 15]$, 200 bootstrap replications, 5 types with covariates at unit-level for the conditional analysis.

Analysis	Confidence intervals	*beta* method	*np* method
unconditional	no	3.2	1.3
unconditional	yes	374.2	176.9
conditional	yes	1870.8	906.8

As highlighted by Table 2, the number n of units has a minor impact, mainly through the preparation stage.

Table 2: Execution time in seconds. Setting: unit-level datasets, $\mathcal{K} = [5, 15]$, options `independencekp` and `noci` for *np* and *beta* methods, 50 draws (default) for *ct* method.

Sample size n	*beta* method	*np* method	*ct* method
1,000	0.30	2.39	0.51
10,000	0.34	2.60	0.80
50,000	0.46	2.19	0.88
100,000	0.67	2.68	1.13

The primary determinant of the computation time is the unit sizes: both the number of distinct values of the support $\mathcal{K}$ and the magnitude of K, as shown by Table 3. With *ct* method, the execution time quickly increases with the magnitude of K while the increase is moderate for *np* method and even lighter for *beta* method.[12]

4 Example

We use the command to measure workplace segregation between natives and foreigners in France (see D'Haultfœuille and Rathelot (2017) for details about the context). A large share of workers is employed in small establishments. This section shows the importance of correcting for the small-unit bias, which may lead to erroneous economic conclusions.

The data used is the 2007 *Déclarations Annuelles des Données Sociales* (DADS), French data linking workers to their employer. Data are exhaustive in the private sector

[12] Remark however that the execution times reported in Table 3 include the draws under random allocation for *ct* method whereas estimation only is performed for *np* and *beta* methods.

Table 3: Execution time in seconds. Setting: unit-level datasets, for each $\mathcal{K}$, $n =$ 10,000 (except 9,000 for the first row) – 1,000 units per distinct size, option `noci` for *np* and *beta* methods, 50 draws (default) for *ct* method.

Support $\mathcal{K}$ of K	*beta* method	*np* method	*ct* method
$[1, 9]$	0.28	0.99	0.23
$[10, 19]$	0.31	2.26	0.57
$[20, 29]$	0.26	5.10	2.16
$[30, 39]$	0.31	7.45	6.26
$[40, 49]$	0.36	8.20	15.1
$[50, 59]$	0.42	12.7	30.7
$[60, 69]$	0.51	11.0	56.6
$[70, 79]$	0.59	15.5	93.1
$[80, 89]$	0.70	22.7	150.3
$[90, 99]$	0.81	24.1	232.0
$[100, 109]$	0.93	26.3	332.1

(1.77 million establishments). In the application, we use the 1.04 million establishments that have between 2 and 25 employees. The minority group consists of individuals born outside of France and with the nationality of a country outside Europe. The overall proportion of minority individuals is 4.1% in the sample studied. Figure 1 shows the estimates of workplace segregation by firm size, for the Duncan, the Theil, the Atkinson (with parameter $b = 0.5$), and the Coworker indices. The Gini index does not satisfy the conditions required by the nonparametric method of HR and is thus not displayed (but see Figure 2 in Appendix 7.2 for the graph on the Gini without the nonparametric estimator).

The distinct methods of the package are used: the estimated bounds $\widehat{\underline{\theta}}$ and $\widehat{\overline{\theta}}$ by *np* method on θ ("np bounds"); the 95%-level confidence interval for this parameter using the modified bootstrap procedure of *np* method, with the default 200 bootstrap iterations ("np CI"); the point-estimate $\widehat{\theta}$ by *beta* method ("beta"); the naive index θ_N ("naive"); the CT-corrected index θ_{CT} using *ct* method with the default 50 draws under random allocation ("ct").

Figure 1 shows that the naive indices overestimate the actual level of segregation: they are almost always above the confidence interval obtained by *np* method (except for the Atkinson index with $K \in \{7, 8\}$). This bias decreases with the size of the units. For the Duncan, the Theil, and the Atkinson indices, the estimated identification interval for θ quickly becomes informative for $K \geq 5$ and reduces to a singleton for $K \geq 9$ (see discussion in Section 2.2). The unit size being larger than 1, the estimated bounds of *np* methods boil down to a point-estimate for the Coworker index.

The point-estimate $\widehat{\theta}$ using *beta* method is within the identification bounds of HR for the Duncan, the Theil, and the Atkinson indices, but is below HR's confidence intervals for the Coworker index. The CT-corrected measure θ_{CT} underestimates the Duncan and Theil indices, being always below the *np* method's confidence interval. θ_{CT} lies

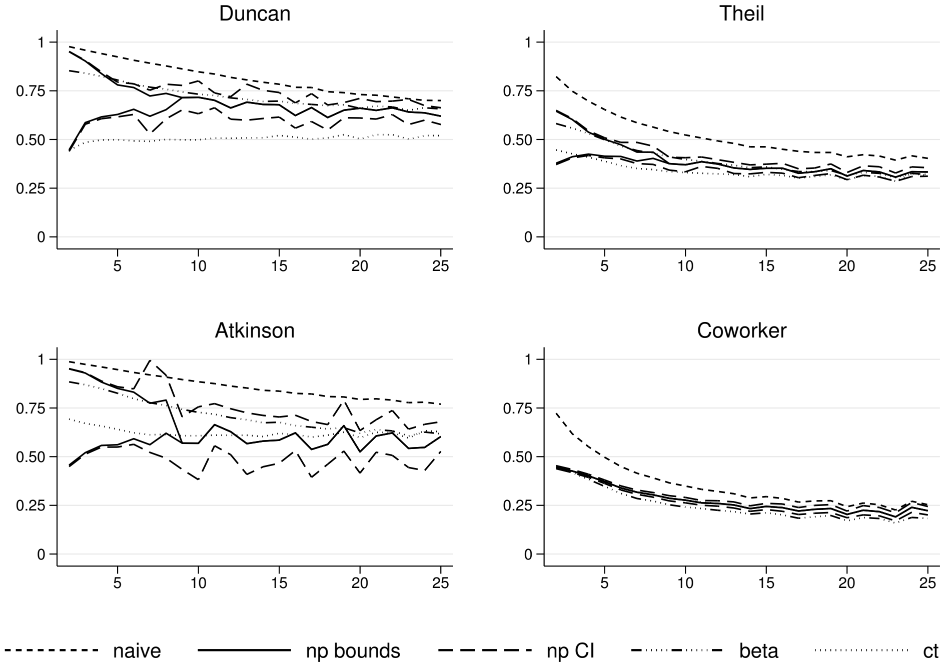

Figure 1: Duncan, Theil, Atkinson, and Coworker indices by firm size.

within the confidence interval and even is quite close to the estimated identification set of θ for the Atkinson and Coworker indices.

Interestingly, the naive indices exhibit a stronger negative relationship between segregation levels and unit size than corrected ones. Neglecting the small-unit bias would produce a statistical artifact as the magnitude of the bias decreases with K and therefore would support a negative correlation while it may not be so. On the contrary, the distribution-based indices that account for the small-unit bias are able to address this question (see Section 5 of HR for further details).

Finally, we report below the Stata output obtained with the `segregsmall` command for *np* method and with option `testbinomial`. Appendix 7.2 displays the output associated with *beta* and *ct* methods. Compared to the analyses of Figure 1 (K by K), the estimation is performed over the entire sample of units ($\mathcal{K} = [2, 25]$) in this output without assuming $K \perp\!\!\!\perp p$. As detailed in Section 3.3, the test of the binomial assumption is automatically performed and saved in this configuration; the option only displays the result in the Stata output. In this application, we cannot reject the binomial assumption at any standard level.

```
. segregsmall K X, format(unit) method(np) testbinomial
```

```
*** Construction of relevant databases for the analysis ***
*** Estimation and inference ***
Estimation - current unit size analyzed (out of 24 distinct sizes):
.........+.........+....
Preparation of bootstrap -
Bootstrap - current bootstrap iteration (out of 200):
.........+.........+.........+.........+.........50
.........+.........+.........+.........+.........100
.........+.........+.........+.........+.........150
.........+.........+.........+.........+.........200

Bounds for segregation indices using nonparametric (np) method:
-----------------------------------------------------------------

Unconditional analysis
Number of units studied in the analysis: 1036840
(0 unit with a single individual are excluded from the analysis)
Number of individuals studied: 6178564
Proportion of minority (or reference) group: 4.1e-02
Assumption on dependence between K and p for estimation and inference: none
Inference:  by bootstrap, 200 repetitions

Unconditional segregation indices:
   Index       Weight-level  |  Lower bound    Upper bound    [95% Conf. Interval]
-----------------------------+-------------------------------------------------------
                             |
   Duncan      unit          |     .58677         .82346        .57864   .8292
   Duncan      individual    |     .63061         .74808        .61966   .75742
   Theil       unit          |     .39246         .52092        .38901   .5251
   Theil       individual    |     .37937         .44251        .37558   .44732
   Atkinson    unit          |     .53907         .83164        .52638   .84667
   Atkinson    individual    |     .56948         .73299        .54977   .7537
   Coworker    unit          |     .37032         .37032        .3674    .37325
   Coworker    individual    |     .31356         .31356        .31084   .31629
Test of binomial assumption (H0: conditional binomial distribution):
(distribution under the null obtained by bootstrap, 200 repetitions)
           Result            |    value of test statistic       p-value
-----------------------------+------------------------------------------
                             |
                             |          1.5398598               .23
```

5 Conclusion

This paper presented the Stata `segregsmall` command which implements three methods (D'Haultfœuille and Rathelot (2017), Rathelot (2012), and Carrington and Troske (1997)) to measure segregation indices in settings when units (neighborhoods, firms, classrooms, etc.) contain few individuals. In such situations, naive indices overestimate the actual level of segregation and produce measures that are not comparable across settings or over time, since the small-unit bias might vary. `segregsmall` enables social scientists to compute segregation indices in those cases and makes the HR nonparametric approach easy to use. It provides asymptotic confidence intervals for HR and R parameters. For all three methods, conditional indices can be estimated: they account for other covariates (either at unit- or individual/position-level) that may influence the allocation process of individuals into units and therefore measure "net" or "residual" segregation. HR and R methods can be used whatever the unit size to measure segregation as a departure from the relevant benchmark of randomness. Even with large units with above one hundred individuals, the parametric approach of R method remains quite affordable as regards computational requirements, even including inference by bootstrap.

6 References

Allen, R., S. Burgess, R. Davidson, and F. Windmeijer. 2015. More reliable inference for the dissimilarity index of segregation. *The Econometrics Journal* 18(1): 40–66.

Åslund, O., and O. N. Skans. 2009. How to measure segregation conditional on the distribution of covariates. *Journal of Population Economics* 22(4): 971–981.

Boisso, D., K. Hayes, J. Hirschberg, and J. Silber. 1994. Occupational segregation in the multidimensional case: decomposition and tests of significance. *Journal of Econometrics* 61(1): 161–171.

Carrington, W. J., and K. R. Troske. 1997. On measuring segregation in samples with small units. *Journal of Business & Economic Statistics* 15(4): 402–409.

Cortese, C. F., R. F. Falk, and J. K. Cohen. 1976. Further considerations on the methodological analysis of segregation indices. *American Sociological Review* 630–637.

D'Haultfœuille, X., and R. Rathelot. 2017. Measuring segregation on small units: A partial identification analysis. *Quantitative Economics* 8(1): 39–73. http://dx.doi.org/10.3982/QE501.

Imbens, G. W., and C. F. Manski. 2004. Confidence intervals for partially identified parameters. *Econometrica* 72(6): 1845–1857.

Krein, M., and A. Nudelman. 1977. The Markov Moment Problem and Extremal Problems Transl. *Mathematical Monographs* 50.

Rathelot, R. 2012. Measuring Segregation When Units are Small: A Parametric Approach. *Journal of Business & Economic Statistics* 30(4): 546–553. http://www.tandfonline.com/doi/abs/10.1080/07350015.2012.707586.

Stoye, J. 2009. More on confidence intervals for partially identified parameters. *Econometrica* 77(4): 1299–1315.

Winship, C. 1977. A revaluation of indexes of residential segregation. *Social Forces* 55(4): 1058–1066.

7 Appendices

7.1 Expressions of the indices in the parametric approach

We use here the same notation as in Section 2.3. If B is a random variable distributed according to the mixture of Beta distributions characterized by v, we have

$$\mu(v) := \mathrm{E}(B) = \sum_{j=1}^{c} \lambda_j \frac{\alpha_j}{\alpha_j + \beta_j}.$$

Duncan index Let $I(t; a, b) = B(t; a, b)/B(a, b)$ with $B(t; a, b) = \int_0^t u^{a-1}(1-u)^{b-1}\,\mathrm{d}u$ the incomplete beta function. Using $B(a, b+1) = B(a, b)b/(a+b)$ and $I(1-t; a, b) = 1$ $-I(t; b, a)$, we obtain

$$\mathrm{D} = \frac{\mu(v)\sum_{j=1}^{c} \lambda_j I(\mu(v); \alpha_j, \beta_j) - \sum_{j=1}^{c} \lambda_j \alpha_j \left(\alpha_j + \beta_j\right)^{-1} I\left(\mu(v); \alpha_j + 1, \beta_j\right)}{\mu(v)(1-\mu(v))}.$$

Theil index To derive the expression of the Theil index as a function of v, we use that if $B \sim \mathrm{Beta}(\alpha, \beta)$, then $1 - B \sim \mathrm{Beta}(\beta, \alpha)$ and $\mathrm{E}[B\ln(B)] = \alpha(\alpha+\beta)^{-1}\{\psi(\alpha+1) - \psi(\alpha+\beta+1)\}$, with ψ the digamma function. This yields

$$\mathrm{T} = 1 - \frac{\sum_{j=1}^{c} \lambda_j \left\{\frac{\alpha_j}{\alpha_j+\beta_j}\psi(\alpha_j+1) + \frac{\beta_j}{\alpha_j+\beta_j}\psi\left(\beta_j+1\right) - \psi(\alpha_j+\beta_j+1)\right\}}{\mu(v)\ln\{\mu(v)\} + (1-\mu(v))\ln\{1-\mu(v)\}}.$$

Atkinson index Let $\Gamma(t) = \int_0^{+\infty} u^{t-1}\exp(-u)\,\mathrm{d}u$ denote the gamma function. Using that $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ and $\Gamma(t+1) = t\Gamma(t)$, the Atkinson index satisfies, for any $b \in (0, 1)$,

$$\mathrm{A(b)} = 1 - \frac{\mu(v)^{\frac{-b}{1-b}}}{1-\mu(v)} \left\{\sum_{j=1}^{c} \lambda_j \frac{\Gamma(\alpha_j + b)\Gamma(\beta_j + 1 - b)}{\Gamma(\alpha_j)\Gamma(\beta_j)(\alpha_j+\beta_j)}\right\}^{\frac{1}{1-b}}.$$

Coworker index If $B \sim \mathrm{Beta}(\alpha, \beta)$, then $\mathrm{E}[B^2] = \alpha(\alpha+1)/\{(\alpha+\beta+1)(\alpha+\beta)\}$. This implies

$$\mathrm{CW} = \left\{\sum_{j=1}^{c} \lambda_j \frac{\alpha_j(\alpha_j+1)}{(\alpha_j+\beta_j+1)(\alpha_j+\beta_j)} - \mu(v)^2\right\} / \left\{\mu(v) - \mu(v)^2\right\}.$$

Gini index Contrary to the previous indices, there is no closed-form expression for the Gini index under a mixture of Beta distributions for F_p because of the term $\int \{F_p(u)\}^2\,\mathrm{d}u$. This quantity has to be approximated by numerical methods. The

Gini index can only be written as

$$G = \left[1 - \mu(v) - \int_0^1 \left\{ \int_0^u \sum_{j=1}^{c} \lambda_j \frac{t^{\alpha_j - 1}(1-t)^{\beta_j - 1}}{B(\alpha_j, \beta_j)} \, \mathrm{d}t \right\}^2 \mathrm{d}u \right] / \left\{ \mu(v) - \mu(v)^2 \right\}.$$

7.2 Supplementary material for the example

Figure 2 is equivalent to the analysis displayed in Figure 1 for the Gini index. Because the Gini index does not satisfy Assumption 1, only the output of *beta* and *ct* methods are reported: the point-estimate $\widehat{\theta}$ ("beta") and the 95%-level asymptotic confidence intervals obtained by bootstrap ("beta CI") using *beta* method (with the default 200 bootstrap iterations); the naive or direct index θ_N ("naive"); the CT-corrected index θ_{CT} using *ct* method with the default 50 draws under random allocation ("ct").

As with the Duncan, the Theil, the Atkinson, and the Coworker indices, Figure 2 illustrates some points discussed in Section 2 in the particular case of the Gini index. Regarding CT correction, there is no reason why θ_{CT} should be close to θ. For the Gini, the CT-corrected index happens to fall far below the confidence interval for the distribution-based index obtained by *beta* method.

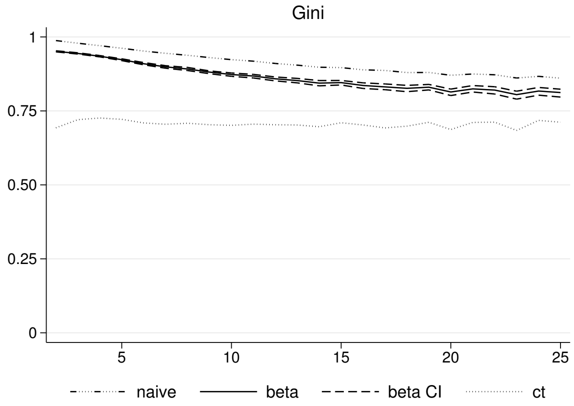

Figure 2: Gini index by firm size.

We report below the Stata output obtained with *beta* and *ct* methods. These estimations are done over the whole sample of units ($\mathcal{K} = [2, 25]$). As an illustration, the option `independencekp` is used for *beta* method.

```
. segregsmall K X, format(unit) method(beta) independencekp repb(400) level(0.98)

*** Construction of relevant databases for the analysis ***

*** Estimation and inference ***
Estimation - K and p assumed independent: units are merged (maximal size = 25)
```

```
Bootstrap - current bootstrap iteration (out of 400):
.........+.........+.........+.........+.........50
.........+.........+.........+.........+.........100
.........+.........+.........+.........+.........150
.........+.........+.........+.........+.........200
.........+.........+.........+.........+.........250
.........+.........+.........+.........+.........300
.........+.........+.........+.........+.........350
.........+.........+.........+.........+.........400

Estimates for segregation indices using parametric (beta) method:
------------------------------------------------------------------

Unconditional analysis
Number of units studied in the analysis: 1036840
(0 unit with a single individual are excluded from the analysis)
Number of individuals studied: 6178564
Proportion of minority (or reference) group: 4.1e-02
Assumption on dependence between K and p for estimation and inference: independence
Inference:  by bootstrap, 400 repetitions

Unconditional segregation indices:
   Index       Weight-level |  Point-estimate    [98% Conf. Interval]
-----------------------------+-----------------------------------------
                             |
   Duncan      unit          |      .75967          .75777   .76129
   Duncan      individual    |      .75967          .75777   .76129
   Theil       unit          |      .43393          .43098   .43639
   Theil       individual    |      .43393          .43098   .43639
   Atkinson    unit          |      .76516          .76254   .76741
   Atkinson    individual    |      .76516          .76254   .76741
   Coworker    unit          |      .2795           .27604   .28258
   Coworker    individual    |      .2795           .27604   .28258
   Gini        unit          |      .89272          .89135   .89388
   Gini        individual    |      .89272          .89135   .89388

. segregsmall K X, format(unit) method(ct) repct(100)

*** Construction of relevant databases for the analysis ***

*** Estimation and correction ***
CT-correction - current random allocation iteration x10 (out of 100):
.........+

Estimates for segregation indices using CT-correction (ct) method:
-------------------------------------------------------------------

Unconditional analysis
Number of units studied in the analysis: 1036840
(0 unit with a single individual are excluded from the analysis)
Number of individuals studied: 6178564
Proportion of minority (or reference) group: 4.1e-02
No inference for naive and CT-corrected indices
CT-correction is made using 100 draws under random allocation (u.r.a.)

Unconditional segregation indices:
   Index       Weight-level |    Naive      Expected u.r.a.     CT-corrected
-----------------------------+------------------------------------------------
                             |
   Duncan           n.a.     |   .85864          .71364              .50634
   Theil            n.a.     |   .57585          .35113              .34632
   Atkinson         n.a.     |   .90735          .75392              .62349
   Coworker         n.a.     |   .41953          .16779              .3025
   Gini             n.a.     |   .94481          .832                .67147
```