Appendix A. Segregation indices as functionals of F

The Gini index

Expression (1) in the main article, relating to the Gini index, is the most difficult to obtain. If $f(.)$ is the pdf of the distribution of the proportions p_a, the Lorenz curve connects the dots of coordinates (x_t, y_t) such that:

$$\begin{aligned} x(p) &= \frac{\int_0^p (1-z)f(z)dz}{1-\mu} \\ y(p) &= \frac{\int_0^p zf(z)dz}{\mu} \end{aligned}$$

The Gini index is defined as twice the area under the curve of the Lorenz curve.

$$G(F) = 2\int_0^1 [x - L(x)]dx = 1 - 2\int_0^1 L(x)dx \text{ with } L(x(p)) = y(p)$$

Substituting $x(p)$ and $y(p)$ by their expressions and changing the variable of integration leads to:

$$\begin{aligned} G(F) &= 1 - 2\int_0^1 L(x(p))x'(p)dp \\ &= 1 - \frac{2}{\mu(1-\mu)}\int_0^1 (1-p)dF(p)\int_0^p zdF(z) \\ &= 1 - \frac{2}{\mu(1-\mu)}\left[\int_0^1 (1-p)F(p)dF(p) - \int_0^1 (1-p)dF(p)\int_0^p (1-z)dF(z)\right] \end{aligned}$$

Using integration by parts,

$$\begin{aligned} \int_0^1 (1-p)F(p)dF(p) &= \frac{1}{2}\int_0^1 F^2(p)dp \\ \int_0^1 (1-p)dF(p)\int_0^p (1-z)dF(z) &= \frac{1}{2}(1-\mu)^2 \end{aligned}$$

Therefore,

$$G(F) = \frac{1}{\mu(1-\mu)}\left(1 - \mu - \int_0^1 F^2(p)dp\right)$$

The dissimilarity index

Expression (2) of the manuscript can be obtained in two ways. It may be deduced by analogy from the expression of $\tilde{D}$, where $\sum_a w_a g(p_a)$ is the empirical counterfactual for $\mathbb{E}[g(p)]$.

It may also be computed directly from his definition: the maximum distance between the Lorenz curve and the diagonal of evenness:

$$D(F) = \max_p x(p) - L(x(p))$$

The FOC of this maximization problem may be expressed as:

$$\frac{(1-p)f(p)}{1-\mu} = \frac{pf(p)}{\mu}$$

This solves for $p = \mu$, which leads to:

$$D(F) = \frac{\int_0^\mu pf(p)dp}{\mu} - \frac{\int_0^\mu (1-p)f(p)dp}{1-\mu}$$

Then, remarking that

$$\int_0^\mu (\mu - p)f(p)dp = \frac{1}{2}\int_0^1 |p-\mu|dF(p)$$

we obtain

$$D(F) = \frac{\int_0^1 |p-\mu|dF(p)}{2\mu(1-\mu)}$$

The Theil index

Expression (3) of the manuscript can easily be obtained by analogy from the expression of $\tilde{H}$, where $\sum_a w_a g(p_a)$ is the empirical counterfactual for $\mathbb{E}[g(p)]$.

Appendix B. The size of the small-unit bias: some simulation results

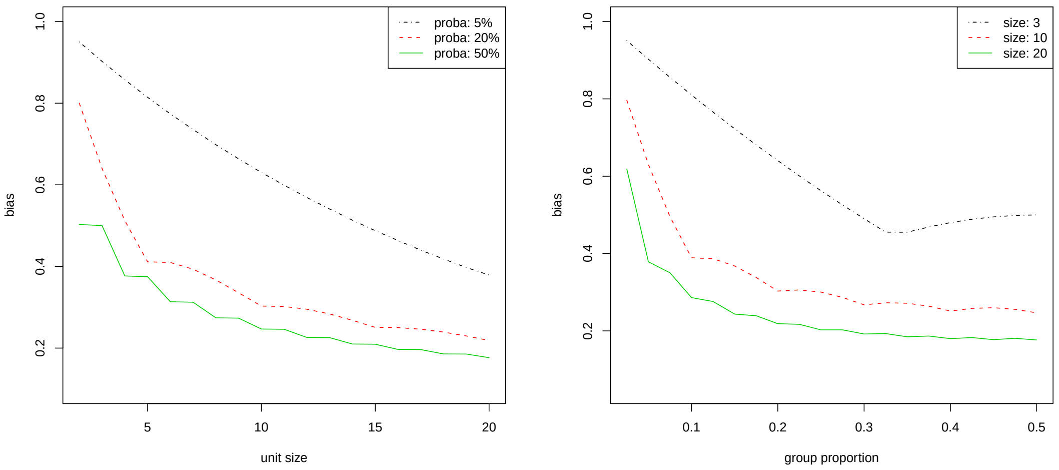
Say that the population is divided into two groups and spread over 10,000 units. The true probability for an individual to belong to the group of interest is assumed to be equal across units: $p_a = \mu, \forall a$. The size of the units is also assumed to be equal in all units: $M_a = M, \forall a$. Each simulation case corresponds to one value of μ and one value of M and, in each case, 1,000 simulations are performed. In a first set of simulations, μ is set to 0.05, 0.2 and 0.5, and M to $2, 3, 4 \ldots 20$. In a second set of simulations, M is set to 3, 10 and 20, and μ to $.025, .05, .075 \ldots .5$. At each step, the p_a are drawn randomly according to the parameters, then N_a are drawn and the dissimilarity index is computed.

Figure 1 plots the mean value of the dissimilarity index in each case, which can be interpreted as bias. Three conclusions need to be drawn from this figure. First, the magnitude of the bias, around 0.5 for units of 15 people and a minority proportion of 5%, makes it an issue one cannot neglect. Then, when the probability is kept constant, the bias decreases with the unit size. Last, when the unit size is kept constant, the bias decreases with the group proportion until this proportion is equal to 50% and increases afterwards. The curve is symmetric around the equiprobability of the groups.

Appendix C. Under which conditions does the CT adjustment lead to $I(F)$?

For simplicity, it is assumed here that all units have the same size M. The sample size A is assumed to be large enough so that the estimation of the expectation of the rv p_a, μ, is assumed not to be an issue : all the analysis here uses μ as a known quantity. The

Figure 1: Mean bias of the dissimilarity index by group proportion and unit size, in the no-segregation case



Source: Simulations by the author.

computations are equivalent for the dissimilarity and the Theil indices, where:

$$h(p) = 1 - \frac{p\log(p) + (1-p)\log(1-p)}{\mu\log(\mu) + (1-\mu)\log(1-\mu)} \text{ for the Theil index } H$$
$$h(p) = \frac{|p-\mu|}{2\mu(1-\mu)} \text{ for the dissimilarity index } D$$

The analysis cannot be directly extended to the case of the Gini index, as it cannot be expressed in the same way. For the Gini, simulation results (see section 4 of the manuscript) are used to measure the distance between the CT adjusted index and $G(F)$.

If $F(.)$ is the cdf of the probability p_a, the index $I(F)$ where $I \in \{H, D\}$ can be written as:

$$I(F) = \int_0^1 h(p)dF(p)$$

while the expectation of the direct index, $\tilde{I}$, is

$$\mathbb{E}(\tilde{I}) = \int_0^1 H_M(p)dF(p)$$

with

$$H_M(p) = \sum_{N=0}^{M} h\left(\frac{N}{M}\right)\left(\begin{array}{c} M \\ N \end{array}\right) p^N(1-p)^{M-N}$$

Note that the expectation of the naive index if there were no segregation, denoted by I^*, is equal to $H_M(\mu)$.

The index corrected by CT method converges to:

$$\mathbb{E}(I_{CT}) = \frac{\mathbb{E}(\tilde{I}) - \hat{I}^*}{1 - \hat{I}^*}$$

Note that, strictly speaking, CT-adjusted indices involve $\hat{I}^*$, simulation-based estimates of I^*, and not directly I^*. However, because $\hat{I}^*$ is arbitrarily close to I^* (it only depends on the number of simulations, not on the unit size or the number of units), we assume that $I^* = \hat{I}^*$ in what follows.

Defining $R_M(p)$ as the rescaled version of $H_M(p)$, this probability limit can be simply written as:

$$\mathbb{E}(I_{CT}) = \int_0^1 R_M(p)dF(p) \text{ with } R_M(p) = \frac{H_M(p) - H_M(\mu)}{1 - H_M(\mu)}$$

The expected difference between the CT adjusted estimator and $I(F)$ is therefore equal to:

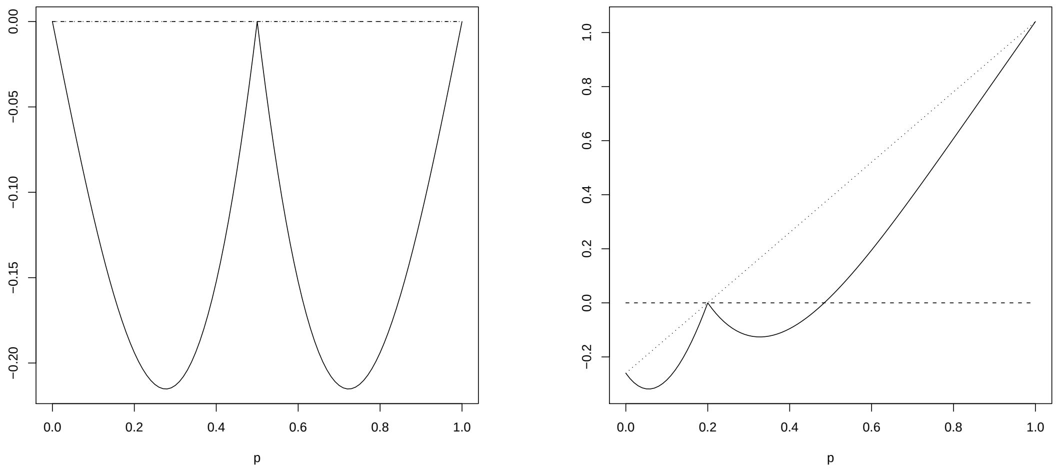
$$\mathbb{E}(I_{CT}) - I(F) = \int_0^1 \left[R_M(p) - h(p)\right] dF(p)$$

The case of the dissimilarity index

Focusing on the dissimilarity index $I = D$, we can precise the expression of the bias. For given M and μ, it is possible to plot $R_M(p) - h(p)$.

The left panel of figure 2 shows the case $M = 5$, $\mu = .5$. In this case, $R_M(p) \leq h(p)$ for all $p \in [0, 1]$. The inequality is strict for all p except $\{0, \mu, 1\}$. Therefore, for all distributions except discrete ones with mass points at 0, 1 or μ, $\mathbb{E}(D_{CT}) < D(F)$.

Figure 2: Dissimilarity index: the function $R_M(p) - h(p)$, $\mu = .5$ (left panel) and $\mu = .2$ (right panel)



Source: Author's computations.

The same result holds when $\mu \neq .5$ but the proof is less trivial. The right panel of figure 2 shows, for instance, the case $M = 5$, $\mu = .2$. In this case, there is a range, when p is large, for which $R_M(p) > h(p)$. However, because the probabilities have expectation .2, this range is less likely than the one around .2, which is globally negative. More formally, computing the maximum of the asymptotic bias across the whole set of the distribution with expectation μ, $\mathcal{D}(\mu)$ and showing that it is non-positive would close the proof. The infinite-dimension problem $\max_{F \in \mathcal{D}(\mu)} \int_0^1 u_M(p) dF(p)$ with $u_M(p) = R_M(p) - h(p)$ reduces to a finite-dimension one, using a result from D'Haultfoeuille and Rathelot (2011). Following the reasoning of their Theorem 2.2, the maximum of the integral w.r.t. F is achieved when F is the cdf of a discrete distribution with at most $M + 1$ mass points.

The problem therefore reduces to:

$$\max_{p_0 \cdots p_M, q_0 \cdots q_M} \sum_{k=0}^{M} q_k u_M(p_k)$$

imposing $\sum_{k=1}^{M} q_k = 1$ and $\sum_{k=1}^{M} q_k p_k = \mu$. The Lagrangian of this problem is:

$$\mathcal{L} = \sum_{k=0}^{M} q_k u_M(p_k) + \lambda_1(1 - \sum_{k=0}^{M} q_k) + \lambda_2(\mu - \sum_{k=0}^{M} q_k p_k)$$

The FOC of the problem lead to, for all $(k, \ell) \in \{0 \dots M\}^2$,

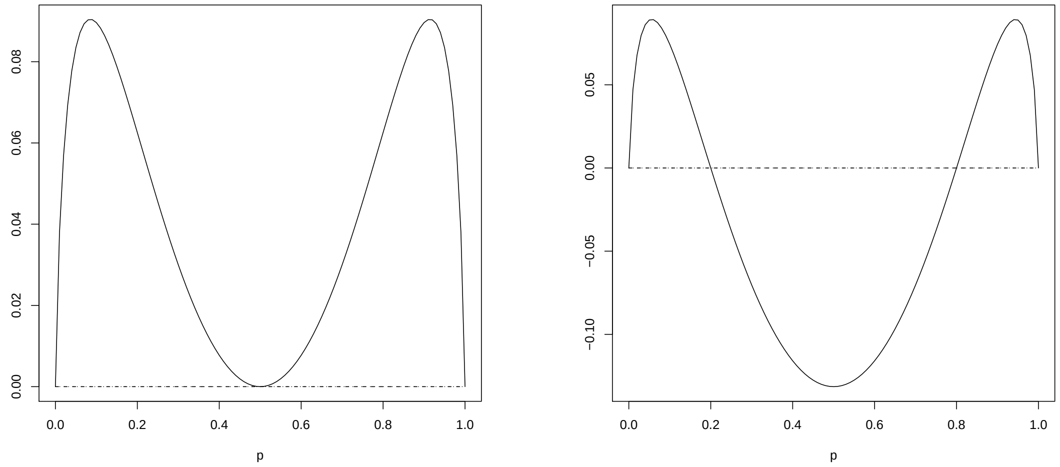
$$u'_M(p_k) = u'_M(p_\ell) = \frac{u_M(p_k) - u_M(p_\ell)}{p_k - p_\ell}$$

The SOC of the problem lead to $u''_M(p_k) < 0, \forall k = 1 \dots M$. Because of the piecewise convexity of the function $u_M(p)$, the SOC do not hold on $]0, \mu[$ and $]\mu, 1[$. Therefore, there is no interior solution for the maximum (while there is one for the minimum). The maximum is obtained on the corner, when the mass points are the extrema $\{0, \mu, 1\}$: this leads to a maximum equal to zero.

The case of the Theil index

For the Theil index, no such reasoning can be done and several discrete and continuous distributions may lead to an unbiased CT-adjusted index. Figure 3 shows, for instance, the function $R_M(p) - h(p)$ in the case $M = 5$, $\mu = .2$. In the extreme case $\mu = .5$, the CT-adjusted index is almost everywhere (except for discrete distributions with mass points $\{0, \mu, 1\}$) above $H(F)$. Otherwise, as soon as $\mu \neq .5$, the function $R_M(p) - h(p)$ is equal to zero for $p = 0, \mu, 1 - \mu, 1$. Around μ, the function may take positive and negative values.

Figure 3: Theil index: the function $R_M(p) - h(p)$, when $\mu = .5$ (left panel) and $\mu = .2$ (right panel)



Source: Author's computations.

Appendix D. Indices as functions of the model parameters

The Lorenz curve (sometimes referred to as the *segregation curve* in the literature) plots the cumulative proportion of the minority population against the cumulative proportion of the majority population after sorting units into ascending order according to the percentage of minority individuals p_a. In the case of a mixture of Beta distributions, the Lorenz curve is defined by the following mapping (see appendix A for a general expression of the Lorenz curve):

$$L(x(p)) = y(p) \text{ with } p \in (0,1)$$

with:

$$x(p) = \frac{1}{1-\mu(v)} \sum_{j=1}^{c} \lambda_j \frac{\beta_j}{\alpha_j+\beta_j} I(p;\alpha_j,\beta_j+1); \quad y(p) = \frac{1}{\mu(v)} \sum_{j=1}^{c} \lambda_j \frac{\alpha_j}{\alpha_j+\beta_j} I(p;\alpha_j+1,\beta_j)$$

where $I(p;\alpha,\beta)$ is the regularized incomplete Beta function, which is also the cdf of the Beta distribution, and $\mu(v) = \sum_j \frac{\lambda_j \alpha_j}{\alpha_j+\beta_j}$ is the expectation of the mixture distribution based on the parameters v.

The expression for the Gini is obtained from its definition, with $L(.)$ the Lorenz curve:

$$G = 2\int_0^1 y - L(y)dy$$

For a given monotone function $g(.)$ such that $y = g(x)$:

$$\begin{aligned} G &= 2\int_0^1 \left[g(x) - L(g(x))\right] g'(x)dx \\ &= 1 - 2\int_0^1 L(g(x))g'(x)dx \end{aligned}$$

Appendix E. Supplementary simulations

Additional simulations, taking the unit size equal to 5 individuals, have also been run. The results, displayed in table 1, show that the conclusions of section 4 of the manuscript remain valid.

Table 1: Simulations: Mean Square Errors with units of 5 individuals

Gini

	Direct	Beta-1	Beta-2	Beta	CT	ABW
B_1	7.53	0.12	0.48	0.15	6.56	3.26
B_2	0.59	0.03	0.04	0.04	1.91	0.19
D_1	3.41	2.08	2.03	2.03	0.04	2.21
D_2	19.65	3.13	0.62	1.46	1.07	11.96
D_3	10.88	0.27	0.60	0.38	6.98	4.89
N	20.41	0.68	1.01	0.73	4.51	10.48
W	8.07	0.12	0.60	0.16	6.87	3.60

Dissimilarity

	Direct	Beta-1	Beta-2	Beta	CT	ABW
B_1	10.75	0.08	0.24	0.10	4.91	5.33
B_2	2.55	0.13	0.06	0.11	2.92	1.16
D_1	10.03	5.49	5.38	5.38	0.10	6.44
D_2	22.08	2.23	0.35	0.91	1.01	13.60
D_3	12.96	0.22	0.34	0.29	4.98	6.47
N	20.97	0.35	0.61	0.39	2.55	11.47
W	11.72	0.08	0.26	0.10	5.02	6.07

Theil

	Direct	Beta-1	Beta-2	Beta	CT	ABW
B_1	7.12	0.03	0.07	0.03	0.10	1.88
B_2	2.64	0.04	0.05	0.05	0.19	0.59
D_1	2.73	0.03	0.03	0.03	0.02	0.79
D_2	8.93	0.12	0.04	0.06	0.04	2.79
D_3	6.71	0.11	0.08	0.12	0.24	1.49
N	9.15	0.03	0.06	0.04	0.04	2.50
W	7.37	0.03	0.06	0.04	0.11	1.99

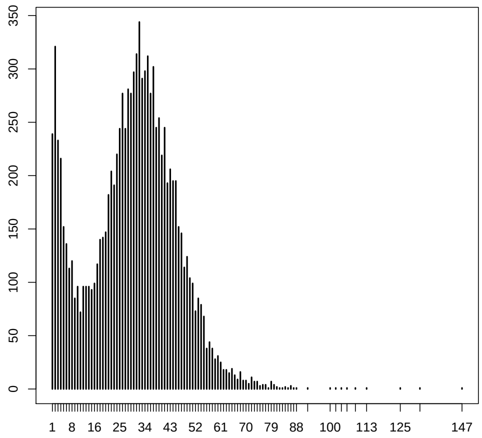
Source: simulations by the author.

Note: For each distribution, simulations are based on 100 draws of samples of 10,000 areal units, each of which with 5 individuals. For the sake of clarity, values in the table are actually 100 times the MSE. For B_1, the dgp is a Beta distribution of parameters (1,9). For B_2, the dgp is a mixture of 2 Beta distribution of parameters (1,9) and (0.1,0.9) with weights (0.3, 0.7). For D_1, the dgp is a discrete distribution of support set (0,0.1,1) with associated weights (0.45,0.5,0.05). For D_2, the dgp is a discrete distribution of support set (0.05,0.1,0.5) with associated weights (0.45,0.5,0.05). For N, the dgp is a truncated normal distribution of mean 0.1 and standard deviation 0.05. For W, the dgp is a truncated Weibull distribution of parameters 0.1 and 1.1.

Appendix F. Application: supplementary material

Figure 4 displays the distribution of the sizes of sampling units. Table 2 provides the Gini, the dissimilarity and the Theil indices computed for the application.

Figure 4: Distribution of the sizes of the sampling units in the French LFS



Source: Labor Force Survey 2005-2008 (Insee).
Note: The y-axis reports the number of units, whether sampling units or sectors, that contain exactly x individuals, x being the figure reported on the x-axis.

Table 2: Segregation indices, by parents' nationalities

	Gini					
	Whole sample		Immigrant sample		French-born sample	
Parents' nationalities	Beta	Direct	Beta	Direct	Beta	Direct
Sub Saharan Africa	0.88 (0.87−0.89)	0.93 (0.93−0.93)	0.91 (0.88−0.93)	0.95 (0.94−0.95)	0.87 (0.84−0.89)	0.96 (0.96−0.96)
North Africa	0.76 (0.75−0.77)	0.82 (0.81−0.83)	0.81 (0.80−0.82)	0.86 (0.86−0.87)	0.70 (0.68−0.71)	0.82 (0.81−0.83)
Middle East	0.84 (0.83−0.85)	0.91 (0.90−0.91)	0.87 (0.86−0.88)	0.92 (0.91−0.92)	0.86 (0.84−0.88)	0.95 (0.94−0.95)
Southern Europe	0.51 (0.50−0.52)	0.63 (0.63−0.64)	0.67 (0.65−0.68)	0.79 (0.78−0.80)	0.47 (0.46−0.49)	0.65 (0.64−0.66)
Northern Europe	0.55 (0.53−0.57)	0.77 (0.77−0.78)	0.76 (0.74−0.78)	0.90 (0.90−0.91)	0.48 (0.43−0.53)	0.81 (0.80−0.81)
Eastern Europe	0.61 (0.60−0.63)	0.81 (0.80−0.81)	0.80 (0.78−0.82)	0.92 (0.92−0.93)	0.58 (0.55−0.60)	0.85 (0.84−0.85)
Asia	0.88 (0.85−0.90)	0.95 (0.95−0.95)	0.94 (0.93−0.95)	0.97 (0.97−0.97)	0.84 (0.80−0.87)	0.97 (0.96−0.97)

	Dissimilarity					
	Whole sample		Immigrant sample		French-born sample	
Parents' nationalities	Beta	Direct	Beta	Direct	Beta	Direct
Sub Saharan Africa	0.75 (0.73−0.76)	0.87 (0.87−0.88)	0.83 (0.81−0.86)	0.91 (0.90−0.91)	0.73 (0.70−0.76)	0.94 (0.94−0.95)
North Africa	0.59 (0.58−0.60)	0.65 (0.65−0.66)	0.66 (0.65−0.67)	0.73 (0.72−0.74)	0.53 (0.52−0.55)	0.68 (0.68−0.69)
Middle East	0.68 (0.67−0.70)	0.81 (0.80−0.82)	0.73 (0.72−0.74)	0.85 (0.84−0.86)	0.72 (0.70−0.75)	0.92 (0.91−0.92)
Southern Europe	0.38 (0.37−0.38)	0.47 (0.46−0.48)	0.50 (0.49−0.51)	0.65 (0.64−0.66)	0.35 (0.34−0.36)	0.49 (0.48−0.50)
Northern Europe	0.40 (0.38−0.42)	0.63 (0.62−0.63)	0.59 (0.56−0.62)	0.84 (0.83−0.85)	0.35 (0.31−0.38)	0.71 (0.70−0.72)
Eastern Europe	0.46 (0.45−0.48)	0.69 (0.68−0.70)	0.64 (0.62−0.67)	0.87 (0.86−0.88)	0.43 (0.41−0.45)	0.77 (0.76−0.78)
Asia	0.80 (0.77−0.83)	0.92 (0.91−0.93)	0.85 (0.83−0.87)	0.95 (0.95−0.96)	0.69 (0.65−0.74)	0.96 (0.95−0.96)

	Theil					
	Whole sample		Immigrant sample		French-born sample	
Parents' nationalities	Beta	Direct	Beta	Direct	Beta	Direct
Sub Saharan Africa	0.31 (0.30−0.33)	0.41 (0.40−0.42)	0.35 (0.34−0.37)	0.43 (0.42−0.44)	0.24 (0.22−0.27)	0.42 (0.41−0.44)
North Africa	0.28 (0.27−0.29)	0.34 (0.33−0.35)	0.30 (0.29−0.31)	0.36 (0.35−0.37)	0.18 (0.17−0.19)	0.29 (0.29−0.30)
Middle East	0.30 (0.29−0.32)	0.39 (0.38−0.40)	0.31 (0.30−0.32)	0.40 (0.39−0.41)	0.26 (0.24−0.27)	0.41 (0.40−0.42)
Southern Europe	0.11 (0.11−0.12)	0.18 (0.18−0.19)	0.17 (0.16−0.18)	0.27 (0.26−0.28)	0.08 (0.08−0.09)	0.18 (0.18−0.19)
Northern Europe	0.12 (0.11−0.13)	0.25 (0.25−0.26)	0.21 (0.19−0.23)	0.36 (0.35−0.37)	0.08 (0.06−0.09)	0.26 (0.25−0.27)
Eastern Europe	0.13 (0.12−0.13)	0.27 (0.26−0.28)	0.23 (0.21−0.25)	0.38 (0.37−0.40)	0.10 (0.09−0.11)	0.29 (0.28−0.30)
Asia	0.30 (0.27−0.32)	0.42 (0.41−0.43)	0.36 (0.34−0.38)	0.47 (0.46−0.48)	0.21 (0.17−0.24)	0.43 (0.43−0.44)

Source: Labor Force Survey 2005-2008 (Insee).

Note: Segregation is measured at the level of the sampling unit of the LFS. The first three columns present the indices computed after the estimation of the Beta model. The last three columns present the indices directly computed with the observed proportions. Confidence intervals at the level of 5% are displayed in parentheses.