Centre de Recherche en Économie et Statistique
CREST

Série des Documents de Travail

n° 2011-06

Measuring Segregation when Units are Small : A Parametric Approach

R. RATHELOT[1]

[1] CREST, 15 boulevard Gabriel Péri, 92245 Malakoff Cedex, France.
Tel. : 33(0)1 41 17 60 36 – Fax : 33(0)1 41 17 60 29. Roland.rathelot@ensae.fr

Measuring Segregation When Units Are Small: A Parametric Approach

Roland Rathelot[§]

Abstract

This paper considers the issue of measuring segregation in a population of units that contain few individuals (*e.g.* establishments, classrooms, ...). When units are small, the usual indices based on sample proportions are biased. We propose a parametric solution: the probability to belong to the minority is assumed to be distributed as a mixture of two beta distributions. The model is estimated and indices are then deduced. Simulations show that this new method performs well compared to existing methods, even in the case of misspecification. An application to residential segregation in France according to parents' nationalities is then proposed.

Keywords: beta-binomial model, mixture of beta distributions, ethnic concentration, dissimilarity index.

[§]CREST, roland.rathelot@ensae.fr, CREST - Bâtiment MK2 Bureau 2020 - Timbre J310 - 15 Boulevard Gabriel Péri - 92245 Malakoff Cedex - France - Tel. : 33 1 41 17 60 36 - Fax. : 33 1 41 17 60 29

1 Introduction

Designing public policies facilitating minorities' social attainment implies understanding their concentration pattern. Knowing how minority groups are spread across schools, firms, neighborhoods is a preliminary step to decision making. There exists an immense number of segregation indices, each with its own properties, which can be used to account for the various dimensions of segregation.[1] However, when units contain few individuals, these indices, because they rely on sample proportions, are poor estimates of the actual level of segregation. Economists and social scientists interested by the distribution of employees across firms, pupils across schools or classrooms, or inhabitants across districts or buildings, may be directly affected by this small-unit bias.[2] The bias is even stronger when the minority group under interest is relatively rare in the total population.

All segregation measures rely on the fact that the observed proportion of the minority group in each unit is a good estimate of the true unobserved probability that a member of this unit belongs to the minority group. When units are small, this assumption may not be realistic. First, assume that the population is made of two equally sized groups (think of men and women, for instance), evenly distributed across units. Given that units have a finite size, observed proportions of each group obviously vary around an average of one half. The smaller the sample in each unit, the higher the variance of these observed proportions are around the true value, one half. This issue is amplified when the population of interest is relatively rare. For instance, let us assume that a minority group represent 5% of total population, that groups are evenly distributed across units, and that, in every units, there are ten observations. Mechanically, it is impossible to observe proportions of one twentieth in these units: only 0, .1, and more rarely .2 or .3, will be observed. Thus, segregation indices are biased when units are small, and this bias increases with the scarcity of the population of interest.

This issue has been first explicited in Cortese, Falk, and Cohen (1976). This text, written

[1] Massey and Denton (1988) attempt to list the dimensions of segregation. More axiomatic approches are also used to categorize segregation indices; see *e.g.* James and Taeuber (1985) or Hutchens (2004).

[2] See Carrington and Troske (1995), Kremer and Maskin (1996), Kramarz, Lollivier, and Pelé (1996), Carrington and Troske (1998a), Carrington and Troske (1998b), Bayard, Hellerstein, Neumark, and Troske (1999) and Hellerstein and Neumark (2008) for examples of studies dealing with workplace segregation. Allen, Burgess, and Windmeijer (2009) or Söderström and Uusitalo (2010) are attempts to account for the small-unit bias in the context of school segregation.

at the climax of the "Index War" that opposed in the late sixties and the seventies several social scientists, mainly sociologists, about the meaning of the dissimilarity index and its appropriateness as a measure of spatial segregation, is the first to mention that it might be sensible to separate *randomness* from *evenness*. What they meant is that the index value should not be compared to zero (the even case) but to the positive value of the index that would be measured if groups were randomly allocated across units. The idea was further explicited in Winship (1977), who proposed an adjusted index that would equal zero in a situation of randomness, instead of evenness. Carrington and Troske (1997) – hereafter CT – develop this idea and introduce a corrected index, close to Winship's one. Their method is the most frequently used in the applied econometrics literature. Allen, Burgess, and Windmeijer (2009) – hereafter ABW – propose to correct indices by a bootstrap procedure and provide simulations that prove its efficiency in removing the small-unit bias. Interestingly, their work emphasizes the issue of inference and proposes statistical tests of segregation.

The main contribution of our paper is to propose a simple parametric approach to remedy the small-unit issue. The approach developed here is based on the estimation of a model generating the probability that an individual in a given unit belongs to the minority group. This probability, which is treated here as a random variable, is assumed to be distributed according to a mixture of two beta distributions whose parameters are to be estimated, taking explicitly into account the fact that smaller units offer less precision. This model is an extension of the beta-binomial model, which was introduced by Skellam (1948) and has been used for decades in contexts as diverse as marketing, education and epidemiology (Lee and Sabavala (1987) propose a short review of this literature). Its main advantage is to offer an appreciable trade-off between flexibility, as the mixture of two beta distributions is appropriate to approximate many kinds of distributions over the $[0, 1]$ segment, and parcimony, as the distribution is summed up by only five parameters. Once the parameters of the distribution are estimated, it is easy to compute concentration indices. Bootstrap is used to provide for inference in this framework.

I also compare the performances of the main existing methods. The CT method is shown to be biased when applied to the dissimilarity index, except when the underlying distri-

bution is discrete with three masspoints – on 0, 1, and the mean of the distribution. For every other distribution, the dissimilarity index corrected with the CT method is found to be below the true value. Simulations are also run in order to compare the methods proposed by CT and ABW to the one that I propose here, for the dissimilarity, the Gini and the Theil indices. These simulations show that the correction method relying on the estimation of the beta mixture performs well in various cases, including those in which the parametric model is misspecified. The CT method frequently overcorrects the dissimilarity and the Gini indices, but achieves satisfying results for the Theil index. The ABW method seems to work well, except when the true level of segregation is low and when the underlying distribution is discrete.

Finally, the beta-mixture correction method is applied to measure the residential segregation of first- and second-generation migrants, according to their country of origin, in the French case. This case illustrates well the small-unit issue. The only available data that allows one to tackle the issue of residential segregation of second-generation migrants is a survey in which the number of individuals by unit is equal to 30 on average. Besides, the groups of interest represent only a few percents of the total population. The gap between directly-computed indices and corrected ones is found to be large and the rank of the indices between groups to be altered by the bias. French individuals whose parents are African immigrants are shown to experience the highest levels of segregation, amongst the non-immigrant groups. Amongst the immigrants, the Asians are the most segregated population.

The next section shows how standard indices behave in the context of small-size units and motivates the relevance of considering alternative approaches. Existing methods which account for the bias are then presented and some of their advantages and drawbacks explicited. Section 3 presents the new method introduced in this paper, how to estimate the distribution parameters, how to derive concentration indices, and how to infer confidence intervals for each quantity. Section 4 displays the performance of the proposed approach on simulated data, whether the model is correctly specified or not. A comparison with other methods is also provided. Section 5 presents an application to the issue of ethnic

residential segregation in the French case. Segregation indices according to parents' origin are computed, using sampling units of the French Labor Force Surveys.

2 The problem and its existing solutions

As surveyed in the widely cited contributions by James and Taeuber (1985) or Massey and Denton (1988), concentration indices are useful tools to capture the unevenness of the distribution of different groups in different units. These units may be, in practice, occupations, industries, firms, schools as well as neighborhoods or metropolitan areas. Populations may be defined according to gender, nationality, ethnicity, social status... In the case of gender, one may expect that the proportions of each group in units oscillate around one half. In the case of nationality, proportions close to zero frequently occur.

2.1 Statistical framework

A population is assumed to be split into two groups, a group of interest and the rest of the population, and to be geographically distributed across A units $a = 1 \dots A$. In the present analysis, the number of individuals in unit a, denoted as M_a, are drawn from a given, unknown distribution. The number of units with M individuals is denoted A_M.

Now assume that there exists a random variable p taking values in $(0, 1)$, that represents the probability for an individual of a given unit to be a member of the population of interest. The cdf of p is denoted by $F_p(.)$. p_a are realizations of this random variable. Unfortunately, while M_a are perfectly observed, p_a are not. What is observed, in each unit, is the realization N_a of a $Binomial(M_a, p_a)$: the number of individuals in unit a that belong to the group of interest. The quantity that is usually used to estimate p_a is $\pi_a = N_a/M_a$, the observed sample proportion of the population of interest in unit a. Note that π_a is unbiased and that, when the number of individuals in unit a M_a goes to infinity, π_a is a consistent estimator of p_a.

This work is focused on the measure of segregation of the population of interest across units. To do so, many indices may be used in practice. For the sake of concision, only three of them, amongst the most frequent ones, are considered in this work: the Gini, the dissimilarity (or Duncan) and the Theil index. The notation I is used to refer to any of

them. A segregation index is a real-valued functional from the space of the distributions defined on the interval $[0, 1]$. Most indices are defined to that they take value on the interval $[0, 1]$. The quantity of interest of this work is $I(F_p)$, the index computed on the distribution F_p of rv p.

In most applied cases, the number of units A is large; in what follows, asymptotic values are obtained for A tends to infinity, with M_a fixed. When A is large, $\mathbb{E}(p)$ is identified: the sample mean $\sum_a N_a / \sum_a M_a$ tends to $\mathbb{E}(p)$ when A tends to infinity, regardless of the unit size. In this case of large A, observing the true probabilities p_a would be a sufficient condition to obtain a fairly good estimate of $I(F_p)$. The issue is that, whenever M_a are small, π_a are going to be poor estimates of p_a and direct indices poor estimates $I(F_p)$. In particular, previous works documented that large biases were to be expected when M_a is small and p_a is close to 0 or 1.

2.2 Three indices

Even if most of the analysis presented in this work can be carried out with any concentration or segregation index, three of them will be used in what follows: the Gini index, the dissimilarity index, and the Theil index.

The indices can be expressed of functions of F_p and $\mathbb{E}(p)$.

$$
\begin{aligned}
G &= \frac{1 - \mathbb{E}(p) - \int_0^1 F_p^2}{\mathbb{E}(p)(1 - \mathbb{E}(p))} \\
D &= \frac{\int_0^1 |z - \mathbb{E}(p)| dF_p(z)}{2\mathbb{E}(p)(1 - \mathbb{E}(p))} \\
H &= 1 - \frac{\int_0^1 z \log z dF_p(z) + \int_0^1 (1 - z) \log(1 - z) dF_p(z)}{\mathbb{E}(p) \log \mathbb{E}(p) + (1 - \mathbb{E}(p)) \log(1 - \mathbb{E}(p))}
\end{aligned}
$$

The sample versions of these indices are often computed, based on the observations M_a and N_a, and defining $\bar{p} = \sum_a N_a / \sum_a M_a$ as the sample analog of $\mathbb{E}(p)$ and $w_a = M_a / \sum_{a'} M_{a'}$ as the weight of unit a in the sample. These sample versions are referred as "direct" or "naive" in what follows as they ignore the small-unit issue.

$$\tilde{G} = \frac{1}{2\bar{p}(1-\bar{p})} \sum_a \sum_{a'} w_a w_{a'} \bar{p}(1-\bar{p})$$

$$\tilde{D} = \frac{1}{2\bar{p}(1-\bar{p})} \sum_a w_a \left|\pi_a - \bar{p}\right|$$

$$\tilde{H} = 1 - \sum_a w_a \frac{\pi_a \log \pi_a + (1-\pi_a)\log(1-\pi_a)}{\bar{p}\log\bar{p} + (1-\bar{p})\log(1-\bar{p})}$$

2.3 The small-unit bias of segregation indices

Segregation measurement using indices suffers from a series of limitations. Duncan and Duncan (1955) list some of them, without giving much formalization. They mention that small unit size might matter for how the indices values are interpreted. Cortese, Falk, and Cohen (1976) and Cortese, Falk, and Cohen (1978) are the first to pinpoint this particular issue. Their work makes it clear that there exists a difference between *evenness* and *random allocation*. Evenness refers to actual equality: if firms have ten employees and if there are as many men and women in the working population, evenness occurs if there are exactly five men and five women in each firm. Random allocation implies that the probabilities for a given individual to be a woman (or a man) are equal in all firms, even if strict equality is not reached. Indices computed using the direct formulas provide information about the distance to evenness whereas practitioners are more interested in the distance to random allocation. The issue is evoked in similar terms in Winship (1977), which, in line with Cortese, Falk, and Cohen (1976), asserts that randomness is a more relevant reference than evenness. Morgan and Norbury (1981) proposes an extension of Winship's framework to the case in which there are more than two groups involved in the index. More recently, the issue has been revived by econometricians, namely by CT and ABW.

In statistical terms, segregation indices computed from observed proportions are biased when units are of finite size. This is because the observed proportions π_a are used to estimate the true probabilities p_a for an individual in the unit a to belong to the group of interest. The bias exists, but how large is it? 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: $p_a = \bar{p}, \forall a$. This represents the

no-segregation case, that is, the case in which both groups are equally spread across units. The size of the units is another parameter, also assumed equal in all units: $M_a = M, \forall a$. In a first set of simulations, $\bar{p}$ is set to .05, .2 and .5, while M is set to $2, 3, 4 \ldots 20$. In a second set of simulations, M is set to 3, 10 and 20, while M is set to $.025, .05, .075 \ldots .5$. At each step, 1,000 simulations are performed.[3] For each simulation, the whole distribution of the individuals is drawn randomly according to the parameters. Then, the index is computed; in this example, the dissimilarity index.

Figure 1 plots the mean value of the dissimilarity index in each case. As these simulations are made under the assumption that there is no segregation, the value of the index must be interpreted as bias. Three conclusions need to be drawn from this figure.[4]

- First, the magnitude of the bias, around .5 for units of 15 people and a minority proportion of 5%, makes it an issue one cannot neglect.
- When the probability is kept constant, the bias decreases with the unit size.
- 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.

2.4 The main existing methods to account for small-unit bias

2.4.1 The linear correction proposed by Carrington and Troske (1997)

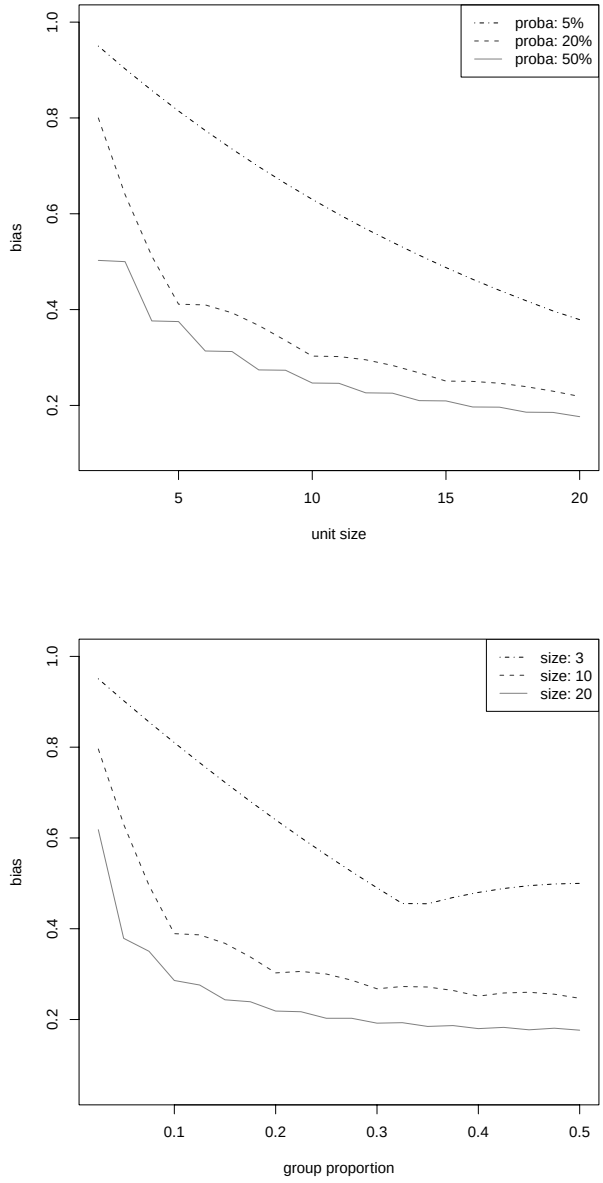
CT propose to adjust the index by subtracting the value of the index obtained under random allocation from the segregation index and to normalize the result so that it ranges from 0 to 1.[5] The method proposed in CT has been used in several applied works, by the authors themselves (Carrington and Troske 1998a), and also, for example, by Hellerstein and Neumark (2008), Dustmann, Glitz, and Schönberg (2009), Persson and Sjögren Lindquist (2010) or Söderström and Uusitalo (2010). Their method has also been extended to the measurement of conditional segregation by Åslund and Skans (2009).

[3] All computations, simulations, estimations, graphical outputs have been made using the statistical software `R`; see R Development Core Team (2010). All programs are available from the author.

[4] The same kind of figure and conclusions can be found in Cortese, Falk, and Cohen (1976) and CT.

[5] Cortese, Falk, and Cohen (1976) proposed a similar method except that the difference is divided by the standard error. Winship (1977) proposed a solution similar to one of CT.

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



Source: Simulations by the author.

The conditions under which the CT correction allows them to get rid of the small-unit bias are not explicited in their paper. In which case is the CT estimator consistent? Let us introduce a few notations. For any index, $\tilde{I}$ refers to the index computed based on a sample of $(N_a, M_a)_{\{a=1...A\}}$; this is the direct or naive measure of segregation. Note that, as a function of the sample values, this index is itself a random variable. As such, one will be interested in the properties of its probability limit, $plim\tilde{I}$, when the number of units A goes to infinity (but not the number of individuals per unit). Conversely, I refers to the unfeasible index, computed using the true p_a, invisible to the practitioner. I is also a rv and one will be interested on the properties of $plimI$. Finally, I^* denotes the random-allocation value of the index. Formally, the random-allocation value is defined as the probability limit of $\tilde{I}$, when all units are assigned the same probability $p_a = \bar{p}$.

The CT corrected index is denoted by:

$$I_{CT} \doteq \frac{\tilde{I} - I^*}{1 - I^*}$$

Most segregation indices take values between zero and one. Zero segregation occurs when, in all units, the probability is equal to the expectation $\bar{p}$. This distribution is denoted as $\mathcal{D}_{\bar{p}}$. In this case, for all samples, $I = plimI = 0$ and $plim\tilde{I} = I^*$. Maximum segregation occurs when the two populations do not cohabit, so that, in some units $p_a = 1$, while in the others, $p_a = 0$. This two-masspoint distribution is denoted as $\mathcal{D}_{0,1}(\bar{p})$, where $\bar{p}$ stands for the expectation of the distribution and in this case the weight on masspoint 1. In this case, for all samples, $I = plimI = \tilde{I} = plim\tilde{I} = 1$.

Let us introduce the distribution defined as the mixture between these two discrete distributions, which plays a key role for the CT correction. In the non-degenerate case, a three-masspoint distribution is obtained, denoted as $\mathcal{D}_{0,\bar{p},1}(w)$, where w is the weight on the distribution $\mathcal{D}_{\bar{p}}$ and $(1-w)$ is the weight on the distribution $\mathcal{D}_{0,1}(\bar{p})$. This distribution is always of expectation $\bar{p}$, and its variance varies from 0, when $w = 1$, to $\bar{p}(1-\bar{p})$, when $w = 0$.

The performance of the method proposed by CT depends on the true distribution of p

and on which index one attempts to compute. Three conclusions can be established for the Theil and the dissimilarity index.[6]

- When the true distribution is one of the family $\mathcal{D}_{0,\bar{p},1}(w)$, the CT correction is exact: $plimI^{CT} = plimI$.

- For the dissimilarity index, this turns out to be an equivalence:

$$plimD_{CT} = plimD \iff \exists(w, \bar{p}), p_a \sim \mathcal{D}_{0,\bar{p},1}(w)$$

  For every other distribution, continuous or discrete, the CT method overcorrects the index: $plimD^{CT} < plimD$.

- For the Theil index, there is no such property. Many distributions may lead to a zero asymptotic bias. Moreover, the asymptotic bias may be positive or negative, depending on the distribution.

Even when an estimator is not consistent, it may lead to results which are sufficiently close to the truth to satisfy the practitioner. The relevant issue is whether, in cases in which researchers are likely to use this correction, the remaining bias after the correction proposed by CT is large or not. This issue is left to section 4 in which simulations are run to compare the performance of the different methods. In a nutshell, CT gives satisfactory results for the Theil index but is severely biased for the dissimilarity and the Gini when the true distribution is not one of the family $\mathcal{D}_{0,\bar{p},1}(w)$.

2.4.2 The bootstrap approach of Allen, Burgess and Windmeijer

ABW use a statistical framework which is similar to the one presented here and allow for the presence of an arbitrary number of groups. Essentially, the authors propose to use bootstrap techniques to adjust the index for the presence of a potential bias. Given the unit size M_a and the observed proportions π_a, they simulate B samples, drawing $N_a(b)$, $b = 1 \dots B$. For each simulated sample $b = 1 \dots B$, an index $I(b)$ (whether the dissimilarity

[6] A formal proof is provided in appendix.

or the Gini index) is computed. The corrected index they propose is then:

$$I_{ABW} \doteq 2\tilde{I} - \frac{1}{B}\sum_{i=1}^{B} I(b).$$

Their idea is that $\tilde{I} - \frac{1}{B}\sum_{i=1}^{B} I(b)$ is an estimator for the small-unit bias and that subtracting it from $\tilde{I}$ provides an estimator for the unbiased estimator. This strategy achieves to reduce the order of the bias from $O(1/M)$ to $O(1/M^{3/2})$ or even $O(1/M^2)$.An additional advantage of their method is that it allows one to perform inference on the indices.In practice, their Monte Carlo simulations show that their correction technique performs well when the unit size or the true segregation level are not too small, that is, when the noise can be separated from the actual unevenness. We investigate this method in more detail in Section 4.

To sum up, even though the existence of a bias has been acknowledged by the literature, there are still few attempts to remedy it. CT propose an easy-to-use corrected index, which proxies well the true index for a restricted family of distributions but may be biased, especially in the case of the dissimilarity index. ABW present an attractive bootstrap-based corrected index, which provides results further from the true value when units are too small or when the true level of segregation is too small. In the next section, an alternative technique, based on parametric assumptions, is presented to correct for small-unit bias.

3 A parametric method

Alternatively to existing methods, this paper proposes a correction method based on a parametric assumption. In a given unit a, the number of minority individuals is assumed to be a $Binomial(M_a, p_a)$. Then, we assume that this probability p_a is a random variable and we specify its distribution: here, a mixture of two beta distributions. In a first step, the parameters of this distribution are estimated, based on the data. In a second step, segregation indices are computed, using the estimated parameters of the distribution of p_a. Inference is performed by full bootstrap on the units.

3.1 Beta distributions and beta mixtures

The support of the distribution of the probability p_a is naturally bounded on the $]0,1[$ segment. Among the different continuous distributions that might suit this requirement, beta distributions are a natural choice. First, as remarked in the early work of Greenwood (1913), a virtue of the beta distribution is that it is the conjugate prior of the binomial distribution, making computations easier. Second, this model is not new: known for decades as the *beta-binomial* model in the statistical literature, it was first formalized by Skellam (1948). In the last half century, it has been then applied in many fields, such as epidemiology, marketing or education sciences, as recalled in Lee and Sabavala (1987).[7] To my knowledge though, it is the first time this kind of model is applied in this context.

Beta distributions are usually chosen both because of their flexibility and their parcimony. Flexible as they encompass many different cases. Examples of beta distributions as well as beta mixtures with expectation fixed to .3, letting the other parameters vary, are featured in figure 2. Parcimonious as the whole distribution is specified by only two parameters. Here, we propose the use of either one or a mixture of two beta distributions. In the latter case, five parameters have to be estimated, but the mixture allows for even more flexibility. Both are tested in the Monte Carlo simulations featured in Section 4.

Formally, the pdf of the rv p_a writes:

$$f_{p_a}(p) \doteq f(p;\alpha_1,\beta_1,\alpha_2,\beta_2,\lambda) = \lambda f_{\alpha_1,\beta_1}(p) + (1-\lambda) f_{\alpha_2,\beta_2}(p)$$

where $f_{\alpha,\beta}(p)$ is the pdf of the beta distribution of parameters α and β.

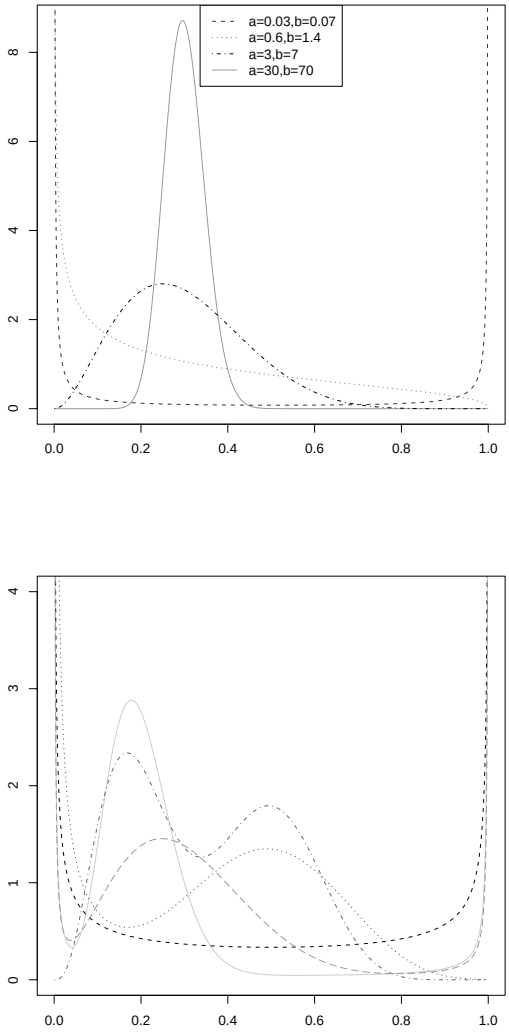
$$f_{\alpha,\beta}(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$$

with $B(.,.)$ the beta function.

The probability that n individuals belong to the group of interest, conditional on p_a and

[7] More recent examples show that the use of this model is rather widespread: Buckley and Schneider (2005) in education sciences, Cogley and Sargent (2009) in macroeconomics, Cox and Katz (1999) in political science or Heitjan (1995) in epidemiology. The beta-logistic model of Heckman and Willis (1977) is an extended version of the beta-binomial model to include covariates.

Figure 2: Examples of beta distributions (top panel) and mixture of two beta distributions (bottom panel), all of expectation .3



Source: Author's computations.

Note: In all these examples, the expectation is fixed at .3.

M_a, is equal to:

$$\mathbb{P}(N_a = n | p_a = p, M_a = m) = \begin{pmatrix} m \\ n \end{pmatrix} p^n (1-p)^{m-n}$$

Integrating this expression with respect to the pdf of the mixture,

$$\mathbb{P}(N_a = n | M_a = m) = \begin{pmatrix} m \\ n \end{pmatrix} \left[\lambda \frac{B(\alpha_1 + n, \beta_1 + m - n)}{B(\alpha_1, \beta_1)} + (1-\lambda) \frac{B(\alpha_2 + n, \beta_2 + m - n)}{B(\alpha_2, \beta_2)} \right] \quad (1)$$

Equation (1) allows one to link the proportions of units of size m in which there are n individuals of the group of interest to the parameters α_1, β_1, α_2, β_2 and λ. The idea of the estimation is to find the five parameters so that the observed proportions are as close as possible to expression (1).

3.2 The estimation

Conditional on the unit size M_a, the probability expressed in (1) is the likelihood that a unit a will contain N_a persons from the population of interest out of a total of M_a. Therefore, summing up over all units with n persons, the log-likelihood may be written:

$$\ell_m(\alpha_1, \beta_1, \alpha_2, \beta_2, \lambda) = \sum_{n=0}^{m} A_m^n \log \left[\lambda \frac{B(\alpha_1 + n, \beta_1 + m - n)}{B(\alpha_1, \beta_1)} + (1-\lambda) \frac{B(\alpha_2 + n, \beta_2 + m - n)}{B(\alpha_2, \beta_2)} \right]$$

where A_m^n is the number of sample units with m individuals, among which n belong to the group of interest.

Then, for each $m > 1$, maximizing $\ell_m(\alpha_1, \beta_1, \alpha_2, \beta_2, \lambda)$ with respect to α_1, β_1, α_2, β_2 and λ provides the estimators $\hat{\alpha}_1(m)$, $\hat{\beta}_1(m)$, $\hat{\alpha}_2(m)$, $\hat{\beta}_2(m)$, $\hat{\lambda}(m)$.[8]

Assuming that the same model holds for a set of units of size belonging to $\mathcal{M} = \{m_1 \dots m_r\}$,

[8] Griffiths (1973) is the first to detail the maximum likelihood estimation of this kind of model. More recent works, like Lee and Sabavala (1987), suggest Bayesian approaches to perform the estimation, especially in the case of small samples or when there exists a prior about the parameters. As, in our application, the sample size is rather large and we have no prior, we estimate the model by MLE.

it is possible to estimate the parameters of the distribution conditional on a set of units and not a single value. In other words, units of different sizes may easily be pooled in the same estimation. The likelihood to be maximized is then:

$$\sum_{m \in \mathcal{M}} \ell_m(\alpha_1, \beta_1, \alpha_2, \beta_2, \lambda)$$

3.3 The concentration indices

Now that the distribution of the probabilities is known, the concentration indices still have to be computed. Two methods may *a priori* be used. The most obvious way to compute the concentration indices associated with each group is to simulate many observations within the mixture of two beta distributions with parameters $\hat{\alpha}$, $\hat{\beta}$, $\hat{\alpha}_2$, $\hat{\beta}_2$, $\hat{\lambda}$, and then to compute the indices using the usual formulae. However, this method may be computationally heavy and can easily be avoided, at least in the case of beta distributions.

Some algebra provide us direct expressions of the indices, as a function of the five parameters. In the case of a mixture of beta distributions, the Lorenz curve (sometimes referred to as the *segregation curve* in the literature) is defined by the following mapping:[9]

$$L(x(z)) = y(z) \text{ with } z \in (0,1) \tag{2}$$

with:

$$\begin{aligned} x(z) &= \lambda \frac{\beta_1}{\alpha_1 + \beta_1} I(z; \alpha_1, \beta_1 + 1) + (1-\lambda) \frac{\beta_2}{\alpha_2 + \beta_2} I(z; \alpha_2, \beta_2 + 1) \\ y(z) &= \lambda \frac{\alpha_1}{\alpha_1 + \beta_1} I(z; \alpha_1 + 1, \beta_1) + (1-\lambda) \frac{\alpha_2}{\alpha_2 + \beta_2} I(z; \alpha_2 + 1, \beta_2) \end{aligned}$$

where $I(z; \alpha, \beta)$ is the regularized incomplete beta function, which is also the cdf of the

[9] To obtain the Lorenz curve, plot the cumulative proportion of the majoritary population as a function of the cumulative proportion of the minoritary population after sorting units into descending order according to the percentage of minoritary individuals p_a. 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:

$$\begin{aligned} x_t &= \frac{\int_0^t (1-z) f(z) dz}{\int_0^1 (1-z) f(z) dz} \\ y_t &= \frac{\int_0^t z f(z) dz}{\int_0^1 z f(z) dz}. \end{aligned}$$

beta distribution. As the dissimilarity index is the maximum distance between the diagonal line and the Lorenz curve, it is straightforward to compute it numerically.

After some computation, the Gini index admits the following expression, using the preceding notations.

$$G(\alpha_1, \beta_1, \alpha_2, \beta_2, \lambda) = 1 - 2 \int_0^1 y(z) \left(\frac{\lambda \beta_1}{\alpha_1 + \beta_1} f(z; \alpha_1, \beta_1 + 1) + \frac{(1-\lambda)\beta_2}{\alpha_2 + \beta_2} f(z; \alpha_2, \beta_2 + 1) \right) dz \tag{3}$$

By definition, the Theil index may be expressed as $1 - \sum E_a / \bar{E}$, where E_a is the entropy associated with each area a and $\bar{E}$, the global entropy, is equal to

$$E_a = -p_a \log p_a - (1 - p_a) \log(1 - p_a) \bar{E} \qquad = -\bar{p} \log \bar{p} - (1 - \bar{p}) \log(1 - \bar{p})$$

Given that p_a is distributed following a $\mathcal{B}(\alpha, \beta)$, the index may be computed directly as a function of the parameters:

$$H(\alpha_1, \beta_1, \alpha_2, \beta_2, \lambda) = 1 - \frac{\lambda E(\alpha_1, \beta_1) + (1 - \lambda) E(\alpha_2, \beta_2)}{\bar{p} \log(\bar{p}) + (1 - \bar{p}) \log(\bar{p})} \tag{4}$$

where

$$\bar{p} = \lambda \frac{\alpha_1}{\alpha_1 + \beta_1} + (1 - \lambda) \frac{\alpha_2}{\alpha_2 + \beta_2}$$
$$E(\alpha, \beta) = \alpha \psi(\alpha + 1) + \beta \psi(\beta + 1) - (\alpha + \beta) \psi(\alpha + \beta + 1)$$

and $\psi(.)$ is the digamma function.

Equations (2), (3) and (4) provide direct expressions that allow one to derive the index of dissimilarity, the Gini index and the Theil index as functions of the estimated parameters of the beta mixture.

Full bootstrap on units is performed in order to provide for inference. At each step, the model is re-estimated, providing new estimates for the segregation indices. Confidence

intervals are computed based on the bootstrap distribution of the estimates. Implicitly, 200 iterations are performed, and 95% confidence intervals are displayed.

4 Simulations

In this section, simulations are used to assess the performance of the method presented in this paper, as well as to compare it to the solutions presented in CT and ABW.

Bias-correcting methods should be robust to the distribution function of p_a as well as to the unit size. As practitioners do not know *a priori* the form of the underlying distribution of the probabilities, they expect these methods to work on the largest possible spectrum of distributions. Beta, beta mixtures, truncated normal, truncated Weibull as well as several discrete distributions are used in the simulations presented in this section. All distributions are calibrated to be of expectation around .1. In the baseline simulations, the unit size is fixed to 10 but, in the appendix, simulations with unit sizes of 3 and 5 are also presented. For each unit size and each distribution, 100 draws in 10,000 units are done. First p_a is drawn i.i.d. in the given distribution. Then, N_a is drawn in a binomial of parameters M_a, p_a.

Results of the simulations are displayed in Table 1 and 2. Table 1 displays average values of the estimates, as well as 95% confidence intervals, while table 2 displays their mean squared errors (MSE). Each panel of the tables is dedicated to a given distribution; the indices are in rows and the methods compared in columns. MSE are computed as the mean of the squared differences between the estimate and the theoretical value of the index (the theoretical value being obtained from the parameters of the distribution).

In table 2, the first column presents the MSE of the direct, or naive, estimates, obtained when one ignores the small-unit bias. These estimates are computed using, as an estimator for probability p_a, the sample proportion π_a. This column gives an idea of the extent of the error that should be expected when direct indices are used. The second and the third columns show the values of the MSE when the parametric method presented in the previous section is used, assuming either a simple beta model (column 2) or a mixture of two beta distributions (column 3). Note that in all but the two first cases, the parametric spec-

Table 1: Simulations: Estimates with units of 10 individuals

	Simulations with a beta model: $\mathcal{B}(1,9)$					
	Unfeasible	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.13 0.13−0.13	0.28 0.27−0.28	0.13 0.12−0.14	0.13 0.12−0.14	0.12 0.11−0.13	0.19 0.18−0.19
Dissimilarity	0.39 0.38−0.39	0.53 0.52−0.54	0.39 0.38−0.40	0.39 0.37−0.40	0.23 0.21−0.24	0.41 0.40−0.42
Gini	0.53 0.52−0.53	0.71 0.70−0.71	0.53 0.51−0.54	0.53 0.51−0.54	0.34 0.32−0.35	0.61 0.60−0.62
	Simulations with mixture of two betas: $.3\mathcal{B}(1,9)+.7\mathcal{B}(.1,.9)$					
	Unfeasible	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.43 0.42−0.44	0.52 0.51−0.53	0.42 0.41−0.44	0.43 0.42−0.44	0.41 0.40−0.43	0.46 0.45−0.47
Dissimilarity	0.67 0.66−0.67	0.73 0.73−0.74	0.69 0.69−0.70	0.67 0.66−0.68	0.56 0.55−0.58	0.67 0.67−0.68
Gini	0.84 0.84−0.85	0.89 0.89−0.90	0.85 0.84−0.86	0.84 0.84−0.85	0.76 0.74−0.77	0.86 0.86−0.87
	Simulations with a $(0,.1,1)$-discrete model					
	Unfeasible	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.50 0.49−0.51	0.59 0.58−0.60	0.47 0.46−0.49	0.49 0.44−0.51	0.50 0.48−0.52	0.53 0.52−0.55
Dissimilarity	0.50 0.49−0.52	0.70 0.69−0.70	0.72 0.71−0.73	0.56 0.53−0.59	0.50 0.48−0.52	0.61 0.60−0.62
Gini	0.75 0.74−0.76	0.89 0.88−0.90	0.87 0.86−0.88	0.78 0.70−0.80	0.75 0.73−0.77	0.85 0.84−0.86
	Simulations with a $(.05,.1,.5)$-discrete model					
	Unfeasible	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.10 0.10−0.11	0.28 0.27−0.28	0.12 0.11−0.12	0.10 0.10−0.11	0.11 0.10−0.12	0.18 0.17−0.18
Dissimilarity	0.24 0.24−0.25	0.49 0.48−0.50	0.37 0.35−0.38	0.26 0.25−0.28	0.16 0.14−0.17	0.34 0.33−0.36
Gini	0.36 0.35−0.37	0.69 0.68−0.70	0.50 0.49−0.52	0.37 0.34−0.40	0.29 0.27−0.31	0.57 0.56−0.59
	Simulations with a $(0,.05,.1,.15,.2)$-discrete model					
	Unfeasible	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.11 0.11−0.11	0.25 0.24−0.25	0.09 0.08−0.10	0.09 0.09−0.11	0.08 0.08−0.09	0.15 0.15−0.16
Dissimilarity	0.33 0.33−0.34	0.50 0.50−0.51	0.32 0.31−0.33	0.33 0.31−0.35	0.18 0.17−0.20	0.38 0.37−0.39
Gini	0.44 0.44−0.45	0.67 0.66−0.68	0.44 0.43−0.46	0.45 0.43−0.48	0.25 0.24−0.26	0.56 0.55−0.57
	Simulations with a truncated normal model					
	Unfeasible	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.04 0.04−0.04	0.21 0.20−0.21	0.04 0.03−0.04	0.04 0.03−0.04	0.04 0.03−0.04	0.10 0.09−0.10
Dissimilarity	0.21 0.20−0.21	0.43 0.43−0.44	0.20 0.19−0.22	0.20 0.19−0.22	0.07 0.06−0.09	0.28 0.27−0.29
Gini	0.29 0.29−0.29	0.60 0.59−0.61	0.29 0.27−0.31	0.29 0.27−0.31	0.11 0.10−0.13	0.46 0.45−0.47
	Simulations with a truncated Weibull					
	Unfeasible	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.13 0.12−0.13	0.28 0.27−0.29	0.13 0.12−0.13	0.13 0.12−0.13	0.12 0.11−0.12	0.19 0.18−0.19
Dissimilarity	0.38 0.37−0.38	0.53 0.52−0.54	0.38 0.37−0.39	0.38 0.37−0.39	0.22 0.21−0.23	0.41 0.40−0.42
Gini	0.52 0.51−0.52	0.71 0.70−0.71	0.52 0.51−0.53	0.52 0.50−0.53	0.33 0.31−0.34	0.61 0.60−0.62

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 10 individuals. 95% confidence interval are showed in parentheses.

Table 2: Simulations: Mean Square Errors with units of 10 individuals

	Simulations with a beta model: $\mathcal{B}(1,9)$				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	2.24	0.00	0.00	0.01	0.33
Dissimilarity	1.98	0.00	0.00	2.60	0.04
Gini	3.22	0.00	0.01	3.62	0.71
	Simulations with mixture of two betas: $.3\mathcal{B}(1,9) + .7\mathcal{B}(.1,.9)$				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.81	0.01	0.00	0.03	0.11
Dissimilarity	0.43	0.07	0.00	1.10	0.01
Gini	0.23	0.00	0.00	0.78	0.04
	Simulations with a $(0,.1,1)$-discrete model				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.81	0.08	0.04	0.01	0.11
Dissimilarity	3.64	4.64	0.33	0.01	1.08
Gini	1.93	1.47	0.12	0.01	0.94
	Simulations with a $(.05,.1,.5)$-discrete model				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	2.96	0.02	0.00	0.01	0.52
Dissimilarity	6.16	1.54	0.05	0.77	1.02
Gini	10.97	2.08	0.04	0.49	4.72
	Simulations with a $(0,.05,.1,.15,.2)$-discrete model				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	1.94	0.04	0.03	0.07	0.18
Dissimilarity	2.83	0.02	0.02	2.27	0.19
Gini	4.99	0.01	0.02	3.76	1.28
	Simulations with a truncated normal model				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	2.80	0.00	0.00	0.00	0.36
Dissimilarity	5.18	0.01	0.01	1.74	0.57
Gini	9.78	0.01	0.01	3.13	2.90
	Simulations with a truncated Weibull				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	2.38	0.00	0.00	0.01	0.37
Dissimilarity	2.34	0.01	0.00	2.60	0.10
Gini	3.58	0.01	0.01	3.64	0.85

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 10 individuals. For the sake of clarity, values in the table are actually 100 times the MSE.

ification assumption is violated: the assumed parametric model does not match the true data-generating process. It is indeed interesting to assess how the parametric correction performs in cases of misspecification. In the fourth column, the MSE related to indices corrected using the CT method are reported. Even though the original paper only deals with the dissimilarity index, the same technique is applied here to the Gini and the Theil indices. The last column presents the value of the MSE when indices are corrected with the ABW bootstrap method. In addition, in table 1, the unfeasible estimate is reported, to make comparisons easier.

Two remarks may be drawn from these tables about the bias suffered by the direct indices. First, the magnitude of the bias, which is always upward, is large compared to the true value of the indices, which stresses that it is important to attempt to correct for small-unit bias. Second, in many cases, the error of the naive estimator is higher than the bias of all presented methods, which shows that, despite their imperfections, existing methods are at least hardly ever harmful.

The comparison of the last four columns underlines the advantages and drawbacks of each method. Consistently with what was found in section 2, the CT index systematically under-estimates the true values of the dissimilarity index, except when the true distribution is a $\mathcal{D}_{0,\mathbb{E}(p),1}$. Interestingly, it seems to be also true for the Gini index. In many cases, the downward biases suffered by the CT-corrected Gini and dissimilarity indices are severe, *e.g.* the dissimilarity index is equal to .07 instead of .21 with a truncated normal. Conversely, the CT correction works remarkably well for the Theil index: the CT-corrected Theil index always lies not more than a few points above or below the true value.[10]

The indices corrected by the method by ABW are upward biased in most cases, which is consistent with the idea that this kind of correction removes only the first order bias. The ABW method performs particularly well, for all three indices, in the case of the beta-mixture, a continuous distribution with a high level of segregation. The ABW method experience larger biases with discrete distributions and when the level of true segregation is low, as in the $\mathcal{D}_{.05,.1,.5}$ or the truncated normal case. In the case of the Theil index, the

[10]This result is not surprising following the analysis in appendix; see figure 5.

ABW method is relatively less efficient than the other methods.

The *simple beta* correction, where the probabilities are assumed to be distributed as a beta distribution, is obviously at its best when the data are drawn in a beta distribution. However, its performance is surprisingly satisfying in other cases, even dealing with discrete distributions or distributions that do not look like beta distributions. For all distributions except the discrete ones $(0, .1, 1)$ and $(.05, .1, .5)$, the error is always quite to zero, for each of three indices. In these two exceptions, the Theil index is almost correct but the Gini and the dissimilarity may be quite distant.

The *beta mixture* correction improves on the previous one and offers more flexibility at the price of less parcimony. The performance of this correction is even better in the cases in which the simple beta correction was already good but it is also better in the other cases. The largest error (.33) is experienced for the dissimilarity index with the $\mathcal{D}_{0,.1,1}$ distribution, a value which remains quite small compared to the performance of other methods in some cases. It also provides almost unbiased results in the case of the Theil index.

Table 3: Simulations: which method should be preferred in which case

	Theil	Dissimilarity	Gini
$\mathcal{B}(1, 9)$	Beta, CT	Beta, ABW	Beta
$.3\mathcal{B}(1, 9) + .7\mathcal{B}(.1, .9)$	Beta, CT	Beta, ABW	Beta, ABW
$\mathcal{D}_{0,.1,1}$	CT, Beta	CT	CT
$\mathcal{D}_{.05,.1,.5}$	Beta, CT	Beta	Beta
$\mathcal{D}_{0,.05,.1,.15,.2}$	Beta, CT	Beta	Beta
truncated Normal $(.1, .05^2)$	Beta, CT	Beta	Beta
truncated Weibull $(1.1, .1)$	Beta, CT	Beta, ABW	Beta

Source: simulations by the author.

Note: This table is a summary of table 1. For each distribution and each index, the least-biased method is reported. If other methods provides estimates with a MSE lower that .10, they appear in second (and third if necessary) positions.

Table 3 sums up the results of the simulations. For each dgp and index, the table reports the method that leads to the estimator with the lowest MSE. The methods which provide indices with a MSE lower than .10 are also reported in second position. Because the beta mixture method outperforms the simple beta one, *Beta* in the table relates to the former.

A rapid glance at the table shows that the beta mixture method is the one that gives the least biased estimates in most cases. The only case in which the beta mixture method does not score well is the dissimilarity index with the discrete distributions $\mathcal{D}_{0,.1,1}$. The second part of the appendix presents more simulation results, in the case in which the number of observation per unit is equal to five. The conclusions of this section remain valid in this case.

5 An application to residential segregation by parents' nationalities in France

This section presents an application of the method introduced in the previous section. The objective is to measure the level of ethnic residential segregation of the population living in France. In the US, most segregation indices are computed using the Censuses, often at the scale of the census tract. In France, it is forbidden by law to collect race or ethnicity variables. The best way to proxy ethnicity is to consider individuals' national origins, which involves not only the nationalities of individuals, but also the nationalities and countries of birth of their ancesters. In practice, the most common way to define ethnicity is to use parents' nationality at birth.[11] Unfortunately, while Censuses provide many variables (social and labor situations, education...) at the scale of the neighborhood, parents' nationality remains absent from this file, still for legal reasons. The largest dataset in which parents' nationalities are observed, since 2005, is the Labor Force Survey.

The sample design of the LFS, though complex, defines *ad hoc* neighborhoods. Households are selected through a three-fold geographical cluster sampling. First, using information from the 1999 Census, *primary units* (of several thousands inhabitants) are selected using stratified random sampling. Then, within each of these primary units, at least one *sector*, consisting of between 120 and 240 contiguous households, is defined. Last, six *sampling units* (named "aires") of, on average, 20 contiguous households are constituted within each sector. Households of one given sampling unit are all interviewed during the same week;

[11] This approach is used in many works dealing with discrimination on the labor market, *e.g.* Domingues Dos Santos (2005), Frickey, Primon, Borgogno, and Vollenweider-Andresen (2005), Silberman and Fournier (2008) or Aeberhardt, Fougère, Pouget, and Rathelot (2010). See Meurs, Pailhé, and Simon (2006) for a description of the few solutions available to social scientists to tackle issues relating to second-generation in France before 2003.

they enter and leave the sample on the same quarter. After their last interview, they are replaced by households of another sampling unit belonging to the same sector. The final dataset provides, for each individual, the ID of the sector and of the sampling unit where he lives. Still, the geographical location of the sampling units and the sectors remains unknown, as these IDs are meaningless.

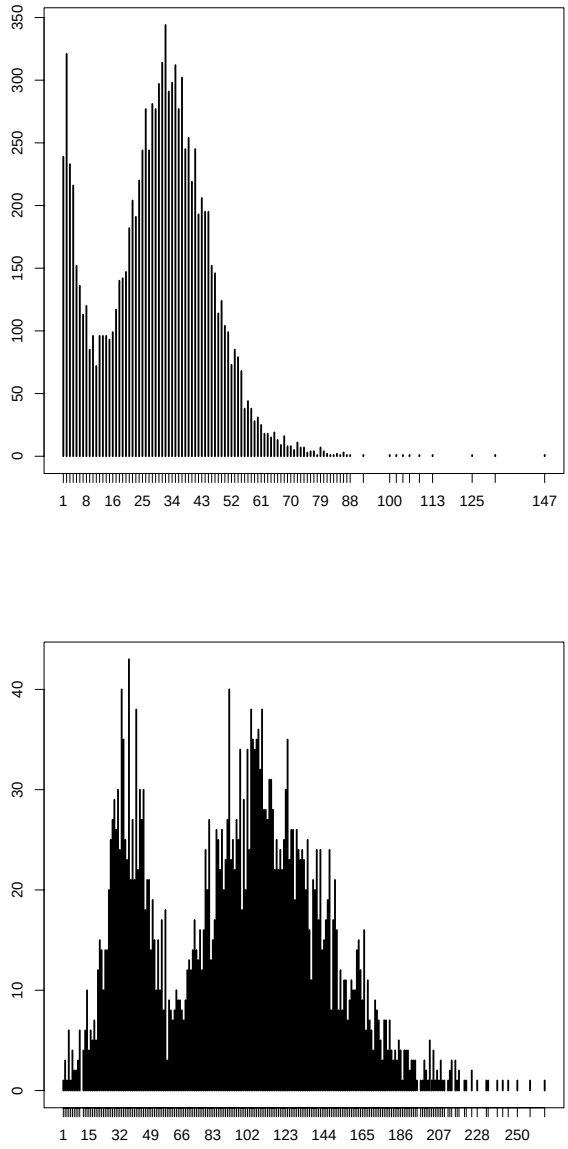
The LFS is likely to be the only dataset allowing to measure ethnic residential segregation in France.[12] However, naive segregation indices would suffer from biases since sampling units are of small size. Figure 3 displays the distribution of the sizes of sampling units and sectors. The mean size of a sampling unit is 30; the median is 31. 25% of the sampling units have a size smaller than 20 and 91% smaller than 50. A sector contains, on average, three to four sampling units. The mean size of a sector is 95; the median is 100. 2% of the sectors have a size smaller than 20 and 25% smaller than 50. Small-unit bias are likely to be an issue when one works with sampling units. For sectors, this issue might be less at stake. Small-unit issues are likely to be aggravated by the fact that minorities are rather rare compared to French individuals of French origin, as reported in the first column of tables 4 and 5. In what follows, we show that if small-unit biases are large for all populations when segregation is computed at the sampling-unit scale, it is also an issue for the rarest populations even at the sector scale.

The LFS from 2005 to 2008 is used to fit a two-beta-distribution mixture model for the proportions of ethnic minorities within sampling units. Three sets of indices are then computed. The first set relates to the total population, whether immigrants or not. In the second set, only individuals of French nationality, born in France or arrived before three years old, are kept in the sample. In the third set, only immigrants arrived in France after the age of 3 are kept. For the sake of simplicity, individuals of this second sample are referred to as French-born, while individuals in the third set is referred to as the immigrants.

The groups are defined according to parents' nationalities. For all populations ("Africa", "Maghreb", "Middle East", "Southern Europe", "Northern Europe", "Eastern Europe" and "Asia"), individuals are required to have at least one parent of the corresponding nation-

[12] Maurin (2004) uses the LFS to obtain concentration measures of social status and ethnicity but, because of the small-unit issue, does not use the usual indices.

Figure 3: Distribution of the sizes of the sampling units (top panel) and the sectors (bottom panel) 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.

ality. To belong to the "France" group, individuals must have both parents born French in France.[13]

Results computed at the scale of sampling units are reported in table 4. Those computed at the scale of sectors are in table 5. The three segregation indices are computed both directly (columns 5 to 7) and following the method presented in this paper (columns 2 to 4). As already noted in the simulations section, there exists large biases between indices computed directly or using this methodology. As a rule of thumb, the bias seems higher when the group is rarer and when the units is smaller (sampling units are on average three to four times smaller than sectors). In our example, populations are not exactly ranked in the same way according to the three indices.

In the whole sample, two groups may be distinguished. Individuals with parents born French or in Europe are the least segregated group. Individuals with parents born with a nationality of Maghreb, Africa, Middle East or Asia are the most segregated group.

Focusing on the group born in France, this ranking is not much changed, the values are smaller but are more contrasted. The least segregated group is the one with French parents. Then comes the group with parents from Southern and Northern Europe. Children of Eastern European migrants are slightly more segregated than those from elsewhere in Europe. Next come the groups with parents from Maghreb and Asia. Finally, the individuals with parents from Middle East and Africa are the most segregated group.

Focusing now on immigrants, indices are substantially higher and closer to each other. The ranking is rather different than in the two former cases. Immigrants from Southern Europe are the least segregated. Next come immigrants from Southern and Eastern Europe and from Maghreb. Immigrants from Africa are slightly more segregated. The most segregated group are the immigrants from Asia.

There exists a small but significant difference between the segregation levels observed at

[13]Note that, according to these definitions, one individual may belong to two groups, if for example, his father came from Algeria and his mother from Poland. In practice, it is only the case for a few tens of people

Table 4: Segregation indices, by parents' nationalities, unit of observation: sampling unit

		Whole sample					
		Beta model			Direct indices		
Parents' nationalities	N	Theil	Dissimilarity	Gini	Theil	Dissimilarity	Gini
France	236234	0.14 (0.13−0.14)	0.35 (0.34−0.35)	0.49 (0.47−0.49)	0.17	0.39	0.54
Africa	2474	0.31 (0.30−0.32)	0.75 (0.73−0.76)	0.88 (0.87−0.89)	0.41	0.87	0.93
Maghreb	14826	0.27 (0.27−0.28)	0.60 (0.59−0.60)	0.76 (0.76−0.77)	0.34	0.65	0.82
Middle East	4462	0.35 (0.29−0.32)	0.75 (0.67−0.71)	0.88 (0.83−0.86)	0.39	0.81	0.91
Southern Europe	18335	0.11 (0.11−0.12)	0.38 (0.37−0.39)	0.52 (0.51−0.53)	0.18	0.47	0.63
Northern Europe	5970	0.12 (0.11−0.13)	0.44 (0.37−0.45)	0.58 (0.52−0.60)	0.25	0.63	0.77
Eastern Europe	4659	0.13 (0.11−0.14)	0.46 (0.44−0.48)	0.61 (0.59−0.63)	0.27	0.69	0.81
Asia	1173	0.28 (0.27−0.30)	0.76 (0.74−0.77)	0.88 (0.87−0.89)	0.42	0.92	0.95
		French-born sample					
		Beta model			Direct indices		
Parents' nationalities	N	Theil	Dissimilarity	Gini	Theil	Dissimilarity	Gini
France	236234	0.14 (0.13−0.14)	0.35 (0.34−0.36)	0.49 (0.48−0.49)	0.17	0.39	0.54
Africa	709	0.24 (0.22−0.26)	0.73 (0.70−0.75)	0.86 (0.84−0.88)	0.42	0.94	0.96
Maghreb	6298	0.19 (0.18−0.19)	0.54 (0.54−0.55)	0.70 (0.69−0.71)	0.29	0.68	0.82
Middle East	1160	0.26 (0.24−0.27)	0.72 (0.71−0.74)	0.86 (0.85−0.87)	0.41	0.92	0.95
Southern Europe	11494	0.08 (0.02−0.09)	0.35 (0.11−0.35)	0.47 (0.15−0.48)	0.18	0.49	0.65
Northern Europe	3582	0.09 (0.08−0.10)	0.39 (0.37−0.41)	0.53 (0.51−0.56)	0.26	0.71	0.81
Eastern Europe	2742	0.10 (0.09−0.11)	0.43 (0.41−0.46)	0.58 (0.56−0.60)	0.29	0.77	0.85
Asia	481	0.20 (0.18−0.22)	0.69 (0.65−0.72)	0.84 (0.80−0.86)	0.43	0.96	0.97
		Immigrants					
		Beta model			Direct indices		
Parents' nationalities	N	Theil	Dissimilarity	Gini	Theil	Dissimilarity	Gini
Africa	1765	0.34 (0.32−0.44)	0.79 (0.77−0.86)	0.91 (0.90−0.94)	0.43	0.91	0.95
Maghreb	8528	0.29 (0.28−0.30)	0.66 (0.65−0.67)	0.81 (0.80−0.82)	0.36	0.73	0.86
Middle East	3302	0.31 (0.30−0.32)	0.73 (0.72−0.74)	0.87 (0.86−0.88)	0.40	0.85	0.92
Southern Europe	6841	0.16 (0.16−0.17)	0.51 (0.50−0.52)	0.67 (0.66−0.68)	0.27	0.65	0.79
Northern Europe	2388	0.21 (0.20−0.23)	0.63 (0.61−0.65)	0.78 (0.76−0.80)	0.36	0.84	0.90
Eastern Europe	1917	0.23 (0.21−0.25)	0.67 (0.64−0.69)	0.82 (0.80−0.83)	0.38	0.87	0.92
Asia	692	0.36 (0.34−0.59)	0.85 (0.83−0.97)	0.94 (0.93−0.99)	0.47	0.95	0.97

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.

Table 5: Segregation indices, by parents' nationalities, unit of observation: sector

		Whole sample					
		Beta model			Direct indices		
Parents' nationalities	N	Theil	Dissimilarity	Gini	Theil	Dissimilarity	Gini
France	236234	0.11 (0.11−0.13)	0.33 (0.32−0.34)	0.45 (0.44−0.47)	0.13	0.33	0.47
Africa	2474	0.24 (0.22−0.25)	0.61 (0.59−0.63)	0.79 (0.78−0.81)	0.29	0.71	0.86
Maghreb	14826	0.22 (0.21−0.23)	0.54 (0.52−0.56)	0.70 (0.69−0.71)	0.25	0.56	0.73
Middle East	4462	0.24 (0.23−0.25)	0.60 (0.59−0.62)	0.77 (0.76−0.78)	0.29	0.67	0.83
Southern Europe	18335	0.09 (0.08−0.09)	0.34 (0.33−0.35)	0.47 (0.45−0.48)	0.11	0.38	0.52
Northern Europe	5970	0.09 (0.08−0.10)	0.34 (0.33−0.36)	0.48 (0.46−0.50)	0.14	0.44	0.61
Eastern Europe	4659	0.08 (0.08−0.10)	0.35 (0.33−0.36)	0.48 (0.47−0.51)	0.15	0.47	0.64
Asia	1173	0.20 (0.18−0.22)	0.61 (0.59−0.64)	0.78 (0.76−0.80)	0.29	0.79	0.88
		French-born sample					
		Beta model			Direct indices		
Parents' nationalities	N	Theil	Dissimilarity	Gini	Theil	Dissimilarity	Gini
France	236234	0.11 (0.11−0.12)	0.33 (0.32−0.33)	0.45 (0.44−0.46)	0.13	0.33	0.47
Africa	709	0.15 (0.13−0.17)	0.56 (0.51−0.60)	0.72 (0.67−0.76)	0.29	0.84	0.90
Maghreb	6298	0.14 (0.13−0.15)	0.47 (0.46−0.49)	0.62 (0.60−0.63)	0.19	0.54	0.70
Middle East	1160	0.19 (0.16−0.21)	0.56 (0.53−0.60)	0.75 (0.70−0.78)	0.29	0.79	0.88
Southern Europe	11494	0.07 (0.06−0.08)	0.32 (0.29−0.33)	0.44 (0.40−0.45)	0.11	0.38	0.52
Northern Europe	3582	0.06 (0.06−0.07)	0.31 (0.29−0.33)	0.43 (0.41−0.46)	0.14	0.47	0.64
Eastern Europe	2742	0.07 (0.06−0.08)	0.33 (0.31−0.36)	0.46 (0.44−0.50)	0.17	0.52	0.69
Asia	481	0.16 (0.12−0.18)	0.61 (0.49−0.68)	0.74 (0.67−0.79)	0.31	0.88	0.92
		Immigrants					
		Beta model			Direct indices		
Parents' nationalities	N	Theil	Dissimilarity	Gini	Theil	Dissimilarity	Gini
Africa	1765	0.27 (0.24−0.28)	0.68 (0.65−0.69)	0.84 (0.82−0.85)	0.32	0.78	0.89
Maghreb	8528	0.24 (0.22−0.25)	0.59 (0.58−0.60)	0.75 (0.74−0.77)	0.27	0.62	0.78
Middle East	3302	0.24 (0.22−0.26)	0.63 (0.62−0.66)	0.80 (0.78−0.81)	0.29	0.71	0.85
Southern Europe	6841	0.11 (0.11−0.12)	0.41 (0.40−0.42)	0.56 (0.54−0.57)	0.16	0.48	0.64
Northern Europe	2388	0.15 (0.13−0.17)	0.48 (0.45−0.50)	0.65 (0.61−0.67)	0.23	0.63	0.79
Eastern Europe	1917	0.15 (0.13−0.17)	0.49 (0.46−0.51)	0.67 (0.63−0.70)	0.24	0.67	0.81
Asia	692	0.26 (0.20−0.26)	0.75 (0.68−0.75)	0.88 (0.83−0.88)	0.34	0.88	0.93

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

Note: Segregation is measured at the level of the sector, a set of sampling units 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.

the sampling-unit scale and the sector scale. Obviously, the larger the unit on which individuals are aggregated, the lower segregation will be. Reardon, Matthews, O'Sullivan, Lee, Firebaugh, Farrell, and Bischoff (2008) show in the case of the US that segregation indices decrease when calculated at a larger scale. Using a index based on a kernel estimation and letting the bandwidth of the kernel vary, Mele (2007) finds, on US data, the same kind of downward relationship.[14] One might object, in the case of sampling units and sectors, that the difference might be due to the difference in unit sizes. For instance, one might claim that, if some bias still remained and if the remaining bias happened to be larger when units were smaller, then, for the same level of true segregation, indices computed on sampling units would indeed be higher than those computed on sectors.

This concern is addressed by computing segregation indices at different scales, fixing the number of individuals per unit. The result of this robustness exercise is reported in table 6. Two samples are built up. For each sector, only one sampling unit is drawn. The first sample is made up with the selected sampling units. The second one is built drawing in each sector the same number of observations as the selected sampling unit. Computing segregation on the first sample gives results at the sampling-unit scale, whereas the second sample provides segregation measures at the sector scale. As the sample size is strictly identical across the two samples, the gap between these measures can be attributed to the scale. The top panel of table 6 should be compared to table 5, and the bottom panel to table 4. If confidence intervals are larger in table 6 compared to those reported in tables 4 and 5, the indices take very similar values. This tends to show that there is indeed a negative relationship between segregation indices and the scale at which they are computed.

[14] Using simulations, he also shows that, if the global trend is always decreasing, there could locally exist, for some particular data generating process, some violations to monotonicity.

Table 6: Segregation indices, by parents' nationalities, immigrants and non-immigrants together

		Random individuals drawn in a sector					
		Beta model			Direct indices		
Parents' nationalities	N	Theil	Dissimilarity	Gini	Theil	Dissimilarity	Gini
France	76380	0.12 (0.12−0.13)	0.33 (0.32−0.35)	0.47 (0.45−0.48)	0.16	0.37	0.52
Africa	810	0.25 (0.22−0.28)	0.68 (0.59−0.70)	0.83 (0.77−0.84)	0.38	0.85	0.91
Maghreb	4681	0.23 (0.22−0.25)	0.55 (0.53−0.57)	0.72 (0.70−0.73)	0.30	0.62	0.79
Middle East	1390	0.25 (0.23−0.27)	0.63 (0.60−0.65)	0.78 (0.76−0.80)	0.36	0.79	0.89
Southern Europe	5913	0.09 (0.08−0.10)	0.34 (0.32−0.35)	0.46 (0.45−0.48)	0.16	0.44	0.59
Northern Europe	1940	0.10 (0.09−0.12)	0.37 (0.34−0.39)	0.51 (0.48−0.55)	0.24	0.61	0.76
Eastern Europe	1447	0.09 (0.08−0.11)	0.36 (0.33−0.40)	0.51 (0.46−0.55)	0.25	0.67	0.78
Asia	381	0.23 (0.19−0.25)	0.69 (0.56−0.72)	0.84 (0.74−0.85)	0.40	0.91	0.94
		One area randomly chosen per sector					
		Beta model			Direct indices		
Parents' nationalities	N	Theil	Dissimilarity	Gini	Theil	Dissimilarity	Gini
France	76334	0.15 (0.14−0.16)	0.36 (0.34−0.37)	0.50 (0.48−0.51)	0.18	0.39	0.55
Africa	794	0.31 (0.25−0.34)	0.75 (0.68−0.76)	0.88 (0.83−0.89)	0.41	0.87	0.93
Maghreb	4816	0.27 (0.26−0.29)	0.60 (0.58−0.61)	0.76 (0.75−0.77)	0.33	0.65	0.82
Middle East	1442	0.30 (0.28−0.33)	0.69 (0.67−0.71)	0.85 (0.82−0.86)	0.39	0.81	0.91
Southern Europe	5888	0.11 (0.10−0.12)	0.36 (0.35−0.38)	0.50 (0.49−0.52)	0.18	0.46	0.62
Northern Europe	1874	0.11 (0.10−0.12)	0.40 (0.38−0.42)	0.55 (0.53−0.57)	0.24	0.62	0.77
Eastern Europe	1493	0.14 (0.12−0.17)	0.43 (0.40−0.47)	0.60 (0.56−0.65)	0.28	0.69	0.81
Asia	372	0.26 (0.20−0.29)	0.73 (0.65−0.76)	0.87 (0.80−0.88)	0.40	0.92	0.94

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

Note: The segregation indices are computed on immigrants and non-immigrants. The sample used for the upper table is built drawing one sampling unit in each sector. The sample used for the bottom table is built drawing in the sector the same number of observations as in the sampling unit drawn for the upper table. 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.

6 Conclusion

When segregation indices are computed using samples in which there are few observations per unit (neighborhoods, school, firms...), they suffer from large upward biases. This is because the proportion of a minority group in a unit is a poor estimate, when this unit is small, of the true probability that an individual of the unit belongs to the minority group. The existence of such a bias hinders any further analysis, such as, for instance, any comparison across groups, metropolitan areas, or countries. In this paper, a new method, parametric and flexible, to compute segregation indices is introduced to deal with these biases. The idea is to assume that the probability that, in a given unit, an individual belongs to the group of interest is a random variable distributed as a mixture of two beta distributions. The parameters of this distribution are estimated and, from these, the values of the segregation indices are deduced. This new method is compared to the two main existing methods, introduced by Carrington and Troske (1997) and Allen, Burgess, and Windmeijer (2009), using simulations. In most cases, which are not restricted to data generating processes distributed as beta mixtures, the new method is showed to perform better than the other two. An application provides the first available figures about ethnic residential segregation in France, using the Labor Force Survey and its unique sampling scheme to define neighborhoods. French individuals whose parents are immigrants experience higher levels of residential segregation than the French whose parents are not immigrants. There are also strong differences across countries of origin: individuals with parents from Subsaharan Africa, Middle East and Northern Africa experience higher levels of residential concentration than those with parents that came from Europe.

Acknowledgments

I thank Romain Aeberhardt, Yann Algan, Mathias André, Elise Coudin, Bruno Crépon, Xavier D'Haultfoeuille, Denis Fougère, Laura Fumagalli, Laurent Gobillon, Albrecht Glitz, Thomas Le Barbanchon, Thierry Magnac, Eric Maurin, Lara Muller, David Neumark, Mirna Safi, Patrick Sillard, Philippe Zamora and participants in the seminars CREST, Insee-D3E, Erudite, and in the ESPE and the Second French Econometrics Conferences for comments and suggestions that have helped improve the paper. Any opinions expressed here are those of the author and not of any institution. All errors remain my own.

References

Aeberhardt, R., D. Fougère, J. Pouget, and R. Rathelot (2010): "Wages and Employment of French Workers with African Origin," *Journal of Population Economics*, 23(3), 881–905.

Allen, R., S. Burgess, and F. Windmeijer (2009): "More Reliable Inference for Segregation Indices," University of Bristol Working Paper No 09/216.

Bayard, K., J. Hellerstein, D. Neumark, and K. Troske (1999): "Why Are Racial and Ethnic Wage Gaps Larger for Men than for Women? Exploring the Role of Segregation Using the New Worker-Establishment Characteristics Database," in *The Creation and Analysis of Employer-Employee Matched Data*, ed. by Haltiwanger, Lane, Spletzer, Theeuwes, and Troske, pp. 175–203. Elsevier Science B.V. (Amsterdam).

Buckley, J., and M. Schneider (2005): "Are Charter School Students Harder to Educate? Evidence from Washington D.C.," *Educational Evaluation and Policy Analysis*, 27(4), 365–380.

Carrington, W. J., and K. R. Troske (1995): "Gender Segregation in Small Firms," *Journal of Human Resources*, 30(3), 503–533.

——— (1997): "On Measuring Segregation in Samples with Small Units," *Journal of Business & Economic Statistics*, 15(4), 402–09.

——— (1998a): "Interfirm Segregation and the Black/White Wage Gap," *Journal of Labor Economics*, 16(2), 231–60.

——— (1998b): "Sex segregation in U.S. manufacturing," *Industrial and Labor Relations Review*, 51(3), 445–464.

Cogley, T., and T. Sargent (2009): "Diverse Beliefs, Survival and the Market Price of Risk," *Economic Journal*, 119(536), 354–376.

Cortese, C., F. Falk, and J. K. Cohen (1976): "Further Considerations on the Methodological Analysis of Segregation Indices," *American Sociological Review*, 41(4), 630–637.

Cortese, C. F., F. Falk, and J. Cohen (1978): "Understanding the Standardized Index of Dissimilarity: Reply to Massey," *American Sociological Review*, 43(4), 590–592.

Cox, G. W., and J. N. Katz (1999): "The Reapportionment Revolution and Bias in U.S. Congressional Elections," *American Journal of Political Science*, 43(3), 812–841.

D'Haultfœuille, X., and R. Rathelot (2011): "Measuring Segregation on Small Units: A Partial Identification Analysis," mimeo CREST.

Domingues Dos Santos, M. (2005): "Travailleurs maghrébins et portugais en France : le poids de l'origine," *Revue Economique*, 56(2), 447–464.

Duncan, O. D., and B. Duncan (1955): "A Methodological Analysis of Segregation Indexes," *American Sociological Review*, 20(2), 210–217.

Dustmann, C., A. Glitz, and U. Schönberg (2009): "Job Search Networks and Ethnic Segregation in the Workplace," mimeo UCL.

Frickey, A., J.-L. Primon, V. Borgogno, and L. Vollenweider-Andresen (2005): *Jeunes diplômés issus de l'immigration: insertion professionnelle ou discriminations*. La Documentation Française.

Greenwood, M. (1913): "On Errors of Random Sampling in Certain Cases not Suitable for the Application of a "Normal" Curve of Frequency," *Biometrika*, 9(1/2), 69–90.

Griffiths, D. A. (1973): "Maximum Likelihood Estimation for the Beta-Binomial Distribution and an Application of the Total Number of Cases of a Disease," *Biometrics*, 29(4), 637–648.

Heckman, J. J., and R. J. Willis (1977): "A Beta-logistic Model for the Analysis of Sequential Labor Force Participation by Married Women," *Journal of Political Economy*, 85(1), 27–58.

Heitjan, D. F. (1995): "Bayesian Analysis of a Protocol for Sampling Preserved Tumor Specimens," *Journal of the American Statistical Association*, 90(429), 38–44.

Hellerstein, J. K., and D. Neumark (2008): "Workplace Segregation in the United States: Race, Ethnicity, and Skill," *The Review of Economics and Statistics*, 90(3), 459–477.

Hutchens, R. (2004): "One Measure of Segregation," *International Economic Review*, 45(2), 555–578.

James, D. R., and K. E. Taeuber (1985): "Measures of Segregation," *Sociological Methodology*, 14, 1–32.

Kramarz, F., S. Lollivier, and L.-P. Pelé (1996): "Wage Inequalities and Firm-Specific Compensation Policies in France," *Annales d'Economie et de Statistique*, 41-42, 369–386.

Kremer, M., and E. Maskin (1996): "Wage Inequality and Segregation by Skill," NBER Working Papers 5718, National Bureau of Economic Research, Inc.

Lee, J. C., and D. J. Sabavala (1987): "Bayesian Estimation and Prediction for the Beta-Binomial Model," *Journal of Business & Economic Statistics*, 5(3), 357–67.

Massey, D. S., and N. A. Denton (1988): "The Dimensions of Residential Segregation," *Social Forces*, 67(2).

Maurin, E. (2004): *Le Ghetto Français*. Seuil.

Mele, A. (2007): "How to Measure Segregation," mimeo Urbana-Champaign.

Meurs, D., A. Pailhé, and P. Simon (2006): "The Persistence of Intergenerational Inequalities linked to Immigration: Labour Market Outcomes for Immigrants and their Descendants in France," *Population*, 61(5), 645–682.

Morgan, B. S., and J. Norbury (1981): "Some Further Observations on the Index of Residential Differentiation," *Demography*, 18(2), 251–256.

Persson, H., and G. Sjögren Lindquist (2010): "The survival and growth of establishments: does gender segregation matter?," in *Research in Labor Economics*, ed. by S. Polachek, and K. Tatsiramos, vol. 30. Emerald Group Publishing Limited.

R Development Core Team (2010): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Reardon, S. F., S. A. Matthews, D. O'Sullivan, B. A. Lee, G. Firebaugh, C. R. Farrell, and K. Bischoff (2008): "The Geographic Scale of Metropolitan Racial Segregation," *Demography*, 45, 489–514.

Silberman, R., and I. Fournier (2008): "Second Generations on the Job Market in France: A Persistent Ethnic Penalty," *Revue Française de Sociologie*, 49(5), 45–94.

Skellam, J. G. (1948): "A Probability Distribution Derived From the Binomial Distribution by Regarding the Probability of Success as Variable Between the Sets of Trials," *Journal of the Royal Statistical Society: Series B*, 10(2), 257–61.

Söderström, M., and R. Uusitalo (2010): "School Choice and Segregation: Evidence from an Admission Reform," *Scandinavian Journal of Economics*, 112.

Winship, C. (1977): "A Revaluation of Indexes of Residential Segregation," *Social Forces*, 55(4), 1058–1066.

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.

Apprendix A. The bias of Carrington and Troske's correction

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, $\bar{p}$, is assumed not to be an issue. All the analysis here uses $\bar{p}$ as a known quantity. The computations are equivalent for the dissimilarity and the Theil indices, where:

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

The analysis driven cannot be directly extended to the case of the Gini index, as it cannot be expressed in the same way. We rely on simulation results (see *infra*) to assess the performance of the CT correction for the Gini.

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

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

The probability limit of the naive index, denoted by $\tilde{I}$, is

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

with

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

Note that the probability limit of the naive index if there were no segregation, denoted by I^*, is equal to $H_M(\bar{p})$.

The index corrected by CT method converges to:

$$plim I^{CT} = \frac{plim \tilde{I} - I^*}{1 - I^*}$$

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

written as:

$$plimI^{CT} = \int_0^1 R_M(p)dF(p) \text{ with } R_M(p) = \frac{H_M(p) - H_M(\bar{p})}{1 - H_M\bar{p}}$$

The asymptotic bias of the CT estimator is therefore equal to:

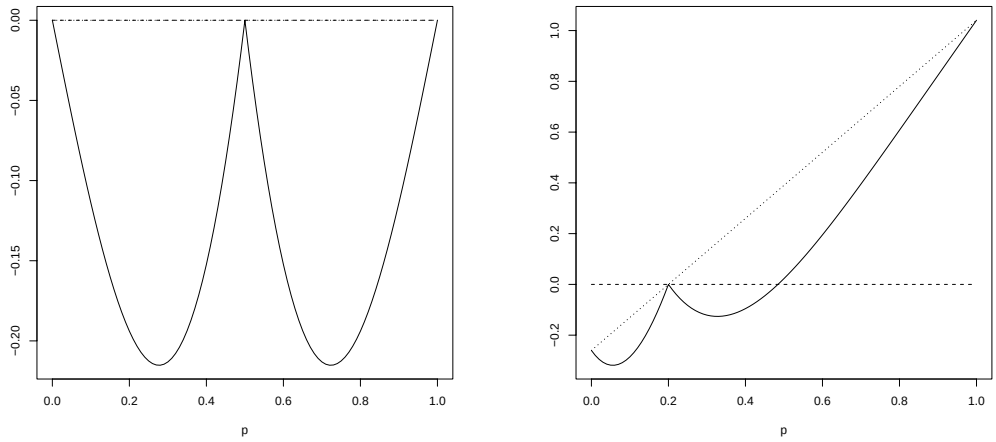
$$plimI^{CT} - plimI = \int_0^1 [R_M(p) - h(p)]\, dF(p)$$

The case of the dissimilarity index

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

The left panel of figure 4 shows the case $M = 5$, $\bar{p} = .5$. In this case, $R_M(p) \leq h(p)$ for all $p \in [0, 1]$. The inequality is even strict for all p except $\{0, \bar{p}, 1\}$. Therefore, for all distributions except discrete ones with masspoints at 0, 1 or $\bar{p}$, $plimD^{CT} < plimD$.

Figure 4: Dissimilarity index: the function $R_M(p) - h(p)$, $\bar{p} = .5$ (left panel) and $\bar{p} = .2$ (right panel)



Source: Author's computations.

The same result holds when $\bar{p} \neq .5$ but the proof is less trivial. The right panel of figure 4 shows, for instance, the case $M = 5$, $\bar{p} = .2$. 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 $\bar{p}$, $\mathcal{D}(\bar{p})$ and showing that it is non-positive would close the proof. The infinite-dimension problem $\max_{F\in\mathcal{D}(\bar{p})} \int_0^1 \delta_M(p)dF(p)$ with $\delta_M(p) = R_M(p) - h(p)$ reduces to a finite dimension one, using a result from D'Haultfœuille and Rathelot (2011). Following the reasoning of Theorem 2.2, the maximum of the integral is achieved when F is the cdf of a discrete distribution with at most $M+1$ masspoints.

The problem therefore reduces to:

$$\max_{p_1 \dots p_M, q_1 \dots q_M} \sum_{k=1}^{M} q_k \delta_M(p_k)$$

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

$$\mathcal{L} = \sum_{k=1}^{M} q_k \delta_M(p_k) + \lambda_1 (1 - \sum_{k=1}^{M} q_k) + \lambda_2 (\bar{p} - \sum_{k=1}^{M} q_k p_k)$$

The FOC of the problem lead to, for all $(k, \ell) \in \{1 \dots M\}^2$,

$$\delta'_M(p_k) = \delta'_M(p_\ell) = \frac{\delta_M(p_k) - \delta_M(p_\ell)}{p_k - p_\ell}$$

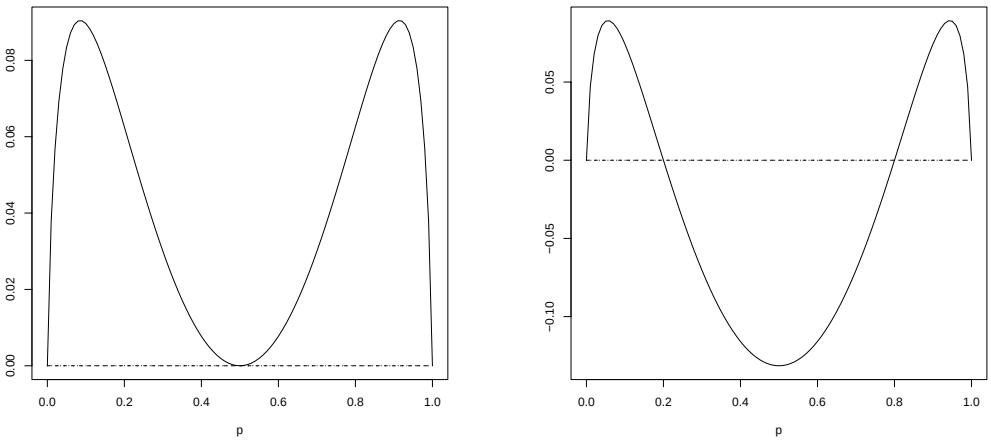
The SOC of the problem lead to $\delta''_M(p_k) < 0, \forall k = 1 \dots M$. There exist only two solutions points that solve the FOC: $p_1^* \in]0, \bar{p}[$ and $p_2^* \in]\bar{p}, 1[$, but the SOC do not hold there. Because of the piecewise convexity of the function $\delta_M(p)$, there is no interior solution for the maximum (while there is one for the minimum). The maximum is obtained when the masspoints are the extrema $\{0, \bar{p}, 1\}$, which all lead 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-corrected index. Figure 5 shows, for instance, the function $R_M(p) - h(p)$ in the case $M = 5$, $\bar{p} = .2$. In the extreme case $\bar{p} = .5$, the CT corrected index is almost everywhere (except for discrete distributions with masspoints $\{0, \bar{p}, 1\}$) upward biased. Otherwise, as soon as $\bar{p} \neq .5$, the function $R_M(p) - h(p)$ is equal

to zero for $p = 0, \bar{p}, 1-\bar{p}, 1$. Around $\bar{p}$, the function may take positive and negative; thus, no systematic bias will undermine the method.

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



Source: Author's computations.

Appendix B. Simulations

In this appendix, we run simulations, taking the unit size equal to 5 individuals. Results, displayed in tables 7 and 8, show that the conclusion of section 4 remain valid.

Table 7: Simulations: Mean Square Errors with units of 5 individuals

	Simulations with a beta model: $\mathcal{B}(1,9)$				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	7.04	0.02	0.01	0.08	1.81
Dissimilarity	10.66	0.09	0.01	4.77	5.24
Gini	7.43	0.17	0.02	6.39	3.17
	Simulations with mixture of two betas: $.3\mathcal{B}(1,9) + .7\mathcal{B}(.1,.9)$				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	2.60	0.01	0.01	0.15	0.56
Dissimilarity	2.53	0.11	0.01	2.80	1.13
Gini	0.58	0.01	0.00	1.80	0.18
	Simulations with a $(0,.1,1)$-discrete model				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	2.65	0.01	0.05	0.01	0.77
Dissimilarity	10.41	5.98	0.45	0.01	6.74
Gini	3.25	2.02	0.10	0.01	2.10
	Simulations with a $(.05,.1,.5)$-discrete model				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	8.98	0.10	0.01	0.01	2.82
Dissimilarity	22.32	2.28	0.14	0.83	13.79
Gini	19.77	3.17	0.19	0.83	12.06
	Simulations with a $(0,.05,.1,.15,.2)$-discrete model				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	6.82	0.34	0.06	0.20	1.56
Dissimilarity	13.22	1.43	0.06	4.64	6.71
Gini	11.03	2.33	0.05	6.44	5.04
	Simulations with a truncated normal model				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	6.79	0.07	0.07	0.20	1.53
Dissimilarity	13.16	0.07	0.07	4.68	6.66
Gini	10.98	0.06	0.05	6.52	5.00
	Simulations with a truncated Weibull				
	Direct	Simple beta	Beta mixture	CT	ABW
Theil	7.44	0.00	0.01	0.07	2.03
Dissimilarity	11.75	0.02	0.01	4.59	6.11
Gini	8.12	0.02	0.02	6.23	3.66

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. 95% confidence interval are showed in parentheses.

Table 8: Simulations: units of 5 individuals

	Simulations with a beta model: $\mathcal{B}(1,9)$					
	True	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.13 0.13−0.13	0.40 0.39−0.40	0.13 0.12−0.14	0.13 0.12−0.15	0.10 0.09−0.11	0.26 0.26−0.27
Dissimilarity	0.39 0.38−0.39	0.71 0.71−0.72	0.38 0.37−0.40	0.39 0.37−0.41	0.17 0.15−0.19	0.62 0.61−0.62
Gini	0.53 0.52−0.53	0.80 0.79−0.80	0.52 0.51−0.55	0.53 0.50−0.55	0.27 0.25−0.30	0.70 0.70−0.71
	Simulations with mixture of two betas: $.3\mathcal{B}(1,9) + .7\mathcal{B}(.1,.9)$					
	True	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.43 0.42−0.44	0.59 0.58−0.60	0.43 0.42−0.44	0.43 0.41−0.45	0.39 0.37−0.41	0.50 0.49−0.51
Dissimilarity	0.67 0.66−0.68	0.83 0.82−0.83	0.70 0.69−0.71	0.67 0.65−0.70	0.50 0.49−0.52	0.78 0.77−0.78
Gini	0.84 0.84−0.85	0.92 0.92−0.92	0.85 0.85−0.86	0.84 0.83−0.86	0.71 0.69−0.73	0.89 0.88−0.89
	Simulations with a $(0,.1,1)$-discrete model					
	True	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.50 0.49−0.52	0.66 0.65−0.68	0.51 0.49−0.52	0.48 0.45−0.51	0.50 0.48−0.52	0.59 0.57−0.60
Dissimilarity	0.51 0.49−0.52	0.83 0.82−0.83	0.75 0.74−0.76	0.57 0.55−0.59	0.50 0.48−0.52	0.76 0.76−0.77
Gini	0.75 0.74−0.76	0.93 0.93−0.94	0.89 0.88−0.90	0.78 0.75−0.80	0.75 0.73−0.77	0.90 0.89−0.90
	Simulations with a $(.05,.1,.5)$-discrete model					
	True	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.10 0.10−0.11	0.40 0.39−0.41	0.13 0.12−0.15	0.11 0.09−0.12	0.11 0.10−0.12	0.27 0.26−0.28
Dissimilarity	0.24 0.24−0.25	0.72 0.71−0.72	0.39 0.37−0.41	0.28 0.25−0.31	0.15 0.13−0.17	0.61 0.60−0.63
Gini	0.36 0.35−0.37	0.80 0.79−0.81	0.53 0.51−0.56	0.39 0.35−0.44	0.27 0.24−0.29	0.70 0.69−0.71
	Simulations with a $(0,.05,.1,.15,.2)$-discrete model					
	True	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.11 0.11−0.11	0.37 0.37−0.38	0.06 0.01−0.10	0.09 0.07−0.10	0.07 0.06−0.07	0.23 0.23−0.24
Dissimilarity	0.33 0.33−0.34	0.70 0.69−0.71	0.25 0.10−0.33	0.31 0.29−0.33	0.12 0.10−0.14	0.59 0.58−0.60
Gini	0.44 0.44−0.45	0.78 0.77−0.78	0.35 0.13−0.46	0.43 0.40−0.46	0.19 0.17−0.22	0.67 0.66−0.68
	Simulations with a truncated normal model					
	True	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.04 0.04−0.04	0.34 0.34−0.35	0.02 0.01−0.04	0.04 0.03−0.08	0.03 0.02−0.04	0.20 0.19−0.20
Dissimilarity	0.21 0.20−0.21	0.67 0.66−0.67	0.14 0.08−0.23	0.21 0.17−0.29	0.05 0.04−0.06	0.55 0.54−0.56
Gini	0.29 0.29−0.29	0.74 0.74−0.75	0.19 0.11−0.31	0.29 0.24−0.40	0.08 0.06−0.11	0.61 0.60−0.62
	Simulations with a truncated Weibull					
	True	Direct	Simple beta	Beta mixture	CT	ABW
Theil	0.13 0.12−0.13	0.40 0.39−0.40	0.13 0.12−0.14	0.13 0.11−0.15	0.10 0.09−0.11	0.27 0.26−0.28
Dissimilarity	0.38 0.37−0.38	0.72 0.72−0.73	0.39 0.37−0.40	0.38 0.35−0.40	0.17 0.15−0.18	0.63 0.62−0.64
Gini	0.52 0.51−0.52	0.80 0.80−0.81	0.53 0.51−0.55	0.52 0.49−0.55	0.27 0.25−0.29	0.71 0.70−0.72

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. 95% confidence interval are showed in parentheses.