More Reliable Inference for the Dissimilarity Index of Segregation

by

Rebecca Allen

Institute of Education, University of London

Simon Burgess

CMPO, University of Bristol

Russell Davidson

McGill University and CIREQ, Montreal
AMSE-GREQAM, Marseille

and

Frank Windmeijer

CMPO, University of Bristol
Cemmap, IFS London

Abstract

The most widely used measure of segregation is the so-called dissimilarity index. It is now well understood that this measure also reflects randomness in the allocation of individuals to units; that is, it measures deviations from evenness, not deviations from randomness. This leads to potentially large values of the segregation index when unit sizes and/or minority proportions are small, even if there is no underlying systematic segregation. Our response to this is to produce adjustments to the index, based on an underlying statistical model. We specify the assignment problem in a very general way, with differences in conditional assignment probabilities underlying the resulting segregation. From this we derive a likelihood ratio test for the presence of any systematic segregation, and bias adjustments to the dissimilarity index. We further develop the asymptotic distribution theory for testing hypotheses concerning the magnitude of the segregation index and show that use of bootstrap methods can improve the size and power properties of test procedures considerably. We illustrate these methods by comparing dissimilarity indices across school districts in England to measure social segregation.

Keywords: segregation, dissimilarity index, bootstrap methods, hypothesis testing

JEL Codes: C12, C13, C15, C46, I21

Correspondence to: Frank Windmeijer, Dept. of Economics. University of Bristol, Bristol BS8 1TN, UK, email: f.windmeijer@bristol.ac.uk, tel: +441179288423

June 2014

1 Introduction

Segregation remains a major topic of research in a number of contexts such as neighbourhoods, workplaces, and schools. Researchers study segregation by poverty status, by sex, and by ethnicity among other characteristics. Almost always, these studies are comparative in some way: for example, arguing that ethnic segregation in neighbourhoods is higher in one city than another, or that segregation by sex in some occupation has changed over time. There is often also an implicit or explicit causal model in mind, and the difference in segregation is associated with some behavioural process. However, the inferential framework for segregation indices is under-developed, a fact that limits the progress that can be made. This paper proposes an approach to strengthen this framework.

It is central to our approach to think of segregation as the outcome of a process of assignment. This includes the assignment of people to neighbourhoods, workers to jobs, or pupils to schools. In general, this allocation is likely to be the result of the interlocking decisions of different agents rather than of a dictator. This perspective offers a number of advantages. First, it ties the outcome to a set of processes that can be analysed and estimated. Second, it makes it clear that the observed outcome is one of a set of possible outcomes, and so naturally leads on to a framework for statistical inference. Third, the connection with the underlying processes makes explicit that it is this systematic or behaviour-based segregation that is the object of interest in terms of analysing the causes of segregation.

There is a large literature concerning the measurement of segregation, with a number of indices in use, all with differing properties. The most widely used measure of segregation is the dissimilarity index, D, defined below (Duncan and Duncan, 1955). It is now widely understood that this measure also reflects randomness in the allocation of individuals to units; that is, it measures deviations from evenness, not systematic segregation. Furthermore, the impact of randomness on D depends on the nature of the context (made precise below). This makes one of the prime tasks in the measurement of segregation difficult, namely making statements on true differences in segregation between cities, school districts, industries, or time periods. For example, the overall proportion of the minority group influences this because a very small minority group is more likely to be unevenly distributed across units by chance, compared to a larger minority group. This problem is particularly acute with small unit sizes. This is easy to see in the following example. Consider a large population, half male and half female. Suppose they are assigned to work in two very large firms. A random assignment process would produce an outcome close to a 50:50 male-female split in each firm and an estimated D of about zero. However, if they were allocated to many firms of size 2, then a random assignment procedure would lead to many all-female firms, many all-male firms and many mixed firms and a high value for D. The high value reflects a strong deviation from evenness despite pure randomness. Others have noted the problem of small unit size in the measurement of segregation, see for instance Carrington and Troske (1997). They proposed an adjustment to segregation indices that has since been used by researchers measuring workplace segregation where small units are particularly likely (*e.g.* Hellerstein and Neumark, 2008) and school segregation (*e.g.* Söderström and Uusitalo, 2005).

In comparing segregation across areas or time, small unit bias should be of concern to researchers for two reasons. First, the size of the bias will differ across comparison areas, potentially leading to an incorrect ranking of levels of segregation across areas. Second, the presence of small unit bias makes a correlation between measured segregation index values and a potentially causal variable, say X, difficult to interpret. It will impact on the estimated effect of X on measured segregation, even if the parameters of the problem (unit size, minority fraction and population) do not vary across areas. In addition, it is likely that the bias as a function of these parameters will be correlated with X, making the true relationship between X and D difficult to identify.

The variable X could for example be income differentials. If one were to investigate racial segregation in schools in an area, one explanation of racial segregation, as indicated by a high value of D, could be income inequality between the two groups. Income inequality could be the cause of neighbourhood and hence school segregation. If there is no income inequality between the two groups in the area, then this could be indicative of behaviour due to other preferences.

In this paper we propose an inferential framework for the canonical segregation measure, D, based on an underlying statistical model. This setup is related to, but different from, that used by Ransom (2000). He derives (asymptotic) inference procedures for D by specifying the sampling variation of a multinomial distribution. We specify the assignment problem in a very general way, and set out the difference in assignment probabilities that underlies the resulting segregation; this is Section 2. From this we derive bias adjustments to D in Section 3, and a likelihood ratio test for the presence of any systematic segregation in Section 4. One of our bias adjustments is based on a simple bootstrap bias correction; other adjustments use the asymptotic normal distribution of the observed frequencies. Following Ransom (2000), we further develop the asymptotic distribution theory for testing hypotheses concerning the magnitude of the segregation index and show that use of bootstrap methods can improve the size and power properties of test procedures considerably; this is in Section 5 and Section 6. As in Ransom (2000), our asymptotic distribution theory relies on the number of units being fixed with unit sizes going to infinity, and our results indicate that our methods work well in settings like our analysis of social segregation in English schools, where the average number of units (schools) in the local authorities are about 55, with the average number of pupils per school equal to 38. Rathelot (2012) recently proposed a Beta-Binomial mixture model to describe segregation and shows that it performs well in a setting with many small units, that is, under asymptotics where the number of units goes to infinity, see also d'Haultfœuille and Rathelot (2011). In Section 7 we present a brief discussion of the measure proposed by Rathelot (2012) and also the one proposed by Carrington and Troske (1997). In Section 8 we illustrate our methods in an example of social segregation in schools in England. Section 9 concludes.

2 Statistical Framework

Underlying an assignment of individuals to units is an allocation process. This might be purely random, or it may be influenced by the actions of agents, including those whose allocation we are studying, as well as others. This systematic allocation process will in general reflect the preferences and constraints of both the individual (such as preferences for racial composition of neighbourhood or ability to pay for houses in a particular neighbourhood) and of the unit to match with particular individuals (such as a firm's desire for highly educated workers or school admissions procedures that favour children of parents of a particular religious denomination). Typically the research question is about characterising segregation arising from this behaviour.

Our notation is as follows. There are units $j = 1, \ldots, J$, nested within an area. Individuals $i = 1, \ldots, n$ either have, or do not have, a characteristic measurable on a dichotomous scale, $c = \{0, 1\}$. This could be black ethnicity, female sex, or poverty status. The number of individuals in the area of status c is denoted n^c, $c = 0, 1$. Individuals are assigned to units and we observe the resulting allocations, n_j^c individuals in unit j having status c. The total number of individuals in unit j is $n_j = n_j^1 + n_j^0$.

There are many indices used to measure segregation (see Duncan and Duncan, (1955), Massey and Denton (1988), and White (1986) for an overview). The formula for each provides an implicit definition of segregation. Massey and Denton characterise segregation along five dimensions: evenness (dissimilarity), exposure (isolation), concentration (the amount of physical space occupied by the minority group), clustering (the extent to which minority neighbourhoods abut one another), and centralisation (proximity to the centre of the city). Throughout this paper we use the index of dissimilarity (denoted D), the most popular unevenness index in the literature. However, our analysis can be extended to other unevenness segregation indices.

The formula for the index of dissimilarity D in the area, which is bounded by 0 (no segregation) and 1, is given by Duncan and Duncan (1955) as [1]

$$D = \frac{1}{2} \sum_{j=1}^{J} \left| \frac{n_j^1}{n^1} - \frac{n_j^0}{n^0} \right|. \tag{1}$$

The basis for an allocation procedure is a set of probabilities p_j^c, which specify the probability that an individual is assigned to unit j, $j = 1, \ldots, J$, conditional on the individual being of status c. We define systematic segregation as being present when there exists j such that $p_j^1 \neq p_j^0$. We can see the relationship between D and the probabilities of the underlying allocation process by noting that the fractions n_j^c / n^c, $c = 0, 1$, are estimates of these probabilities. With $\hat{p}_j^c = n_j^c / n^c$, the index of dissimilarity (1) is just one half of $\sum_{j=1}^{J} |\hat{p}_j^1 - \hat{p}_j^0|$.

Formally the allocation process is as follows. An area population of n individuals, with a given proportion $p = n^1/n$ with status $c = 1$, is allocated to J units according to the probabilities p_j^c. Each individual is allocated independently of the others. All implicit dependencies of group formations are captured by the allocation probabilities p_j^c. The outcomes of this process are the allocations n_j^1 and n_j^0 . Clearly, unit sizes are not fixed in

[1] D measures the share of either group that must be removed, without replacement, to achieve zero segregation (Cortese *et al.*, 1976; Massey and Denton, 1988). It can be shown to be equal to the maximum distance between the line of equality and a segregation curve that sorts units by p_j, then plots the cumulative share of $c = 1$ individuals against the cumulative share of $c = 0$ individuals (Duncan and Duncan, 1955).

this setup as they are equal to $n_j = n_j^1 + n_j^0$ and are therefore determined by the stochastic allocation. The expected unit sizes are given by $\mathrm{E}(n_j) = n^1 p_j^1 + n^0 p_j^0$.
We can now interpret the index of dissimilarity as an estimator for the population quantity

$$D_{\text{pop}} = \frac{1}{2} \sum_{j=1}^{J} |p_j^1 - p_j^0|.$$

It is clear that $D_{\text{pop}} = 0$ if and only if $p_j^1 = p_j^0$ for all $j = 1, \ldots, J$.
From the allocation process described above, we can estimate the conditional probabilities by maximum likelihood. As the allocations are two independent multinomial distributions, the log-likelihood function, given the observed allocations, is given by

$$\log L = \log\Big(\frac{n^1!}{n_1^1! \ldots n_J^1!}\Big) + \log\big(\frac{n^0!}{n_1^0! \ldots n_J^0!}\big) + \sum_{j=1}^{J} n_j^1 \log p_j^1 + \sum_{j=1}^{J} n_j^0 \log p_j^0. \tag{2}$$

Clearly, the maximum likelihood estimates are given by $\hat{p}_j^1 = n_j^1/n^1$ and $\hat{p}_j^0 = n_j^0/n^0$, $j = 1, \ldots, J$, exactly the same as the estimates that appear in D.
Ransom (2000) proposed the following statistical model for a random sample of size n:

$$\Pr(n_1^0, n_2^0, \ldots, n_J^0, n_1^1, n_2^1, \ldots, n_J^1; \pi_{jc}) = n! \prod_{j=1}^{J} \prod_{c=0}^{1} \frac{(\pi_{jc})^{n_j^c}}{n_j^c!},$$

where π_{jc} is the joint probability of observing an individual with status c and in unit j in the sample. Mora and Ruiz-Castillo (2007), and references therein, consider a similar setup for an information index of multi-group segregation. Ransom (2000, p. 458) notes that this model is not appropriate when the population is observed, as then the π_{jc} are known. The parameters π_{jc} are not those that enter the segregation index D_{pop}, which are the conditional probabilities $p_j^c = \Pr(\text{unit} = j \mid c) = \pi_{jc} / \sum_{s=1}^{J} \pi_{sc}$.
Our model is applicable even when we observe the complete, finite population, because randomness is achieved by the random allocation process to units. Our statistical model is for a finite population of fixed size $n = n^0 + n^1$, with parameters p_j^c, $j = 1, \ldots, J$, $c = 0, 1$, and is given by

$$\Pr(n_1^0, n_2^0, \ldots, n_J^0, n_1^1, n_2^1, \ldots, n_J^1; n^0, n^1; p_j^c) = \prod_{c=0}^{1} n^c! \prod_{j=1}^{J} \frac{(p_j^c)^{n_j^c}}{n_j^c!}.$$

The logarithm of this expression is just the log-likelihood (2). In the remainder of the paper we will focus on this particular model.
Our design is particularly well suited for our analysis of social segregation in schools in England. The provision of education in England is organised at the district, or local authority level. Pupils within a local authority have to choose a school within that district with certain limitations due, for example, to catchment area requirements. School (cohort)

sizes vary substantially within a local authority, and school cohort sizes vary over time owing to changing demand and size of the cohort population.

A different model would apply if the unit sizes n_j are assumed fixed, as well as the population size n and minority fraction $p = n^1/n$. In this case, the allocation mechanism is determined by the probability that an individual has status c conditional on being in unit j instead of the other way round. However, no matter whether unit sizes are random or fixed, D is still an estimator of D_{pop} if instead of the full population we obtain a random sample drawn from it.

2.1 Bias

As D is an estimator of D_{pop}, we define the bias of D as $\mathrm{E}(D) - D_{\text{pop}}$, where the expectation is taken over the independent multinomial distributions with probabilities p_j^c, $j = 1, \ldots, J$, $c = 0, 1$. For given population size n and minority proportion p, we have

$$\mathrm{E}(D) = \frac{1}{2} \sum_{\{n_1^0, \ldots, n_J^0\}} \sum_{\{n_1^1, \ldots, n_J^1\}} \left[\left(\sum_{j=1}^{J} \left| \frac{n_j^1}{n^1} - \frac{n_j^0}{n^0} \right| \right) \prod_{c=0}^{1} n^c! \prod_{j=1}^{J} \frac{(p_j^c)^{n_j^c}}{n_j^c!} \right].$$

The value of $\mathrm{E}(D)$ is a function of the underlying conditional probabilities, summarised by D_{pop}, and of unevenness generated by the randomness of the allocation process. As has been well documented in the literature, see for instance Carrington and Troske (1997), D can be severely upward biased when unit sizes are small and there is no systematic segregation, that is, $p_j^1 = p_j^0$ for all j and $D_{\text{pop}} = 0$. Intuitively, this bias arises because D is the sum of the absolute value of differences between the minority and majority proportions in a unit. Suppose that in unit j, $p_j^1 = p_j^0$. Sampling variation in the estimated proportions will almost surely lead to non-zero estimated differences, especially if unit sizes are small. Since the dissimilarity index sums the absolute values of these differences, this will lead to an upward bias in the index.

For a small number of units J and small unit sizes, we can calculate the expected value of D analytically. Figure 1 graphs the bias $\mathrm{E}(D) - D_{\text{pop}}$ for $J = 4$, $n = \{20, 40, 60\}$, $p = 0.1$, and for various values of D_{pop}. These values of D_{pop} are obtained by setting the p_j^c according to a scheme discussed in Section 3 below. The expected unit sizes are the same for the 4 units, *i.e.* 5 when $n = 20$, 10 when $n = 40$, and 15 when $n = 60$. The small-unit bias is apparent in the figure. When expected unit sizes are equal to 5, $\mathrm{E}(D)$ is equal to 0.56 when $D_{\text{pop}} = 0$. The graph also shows that the bias is a decreasing function of increasing systematic segregation (D_{pop}) and a decreasing function of expected unit size.

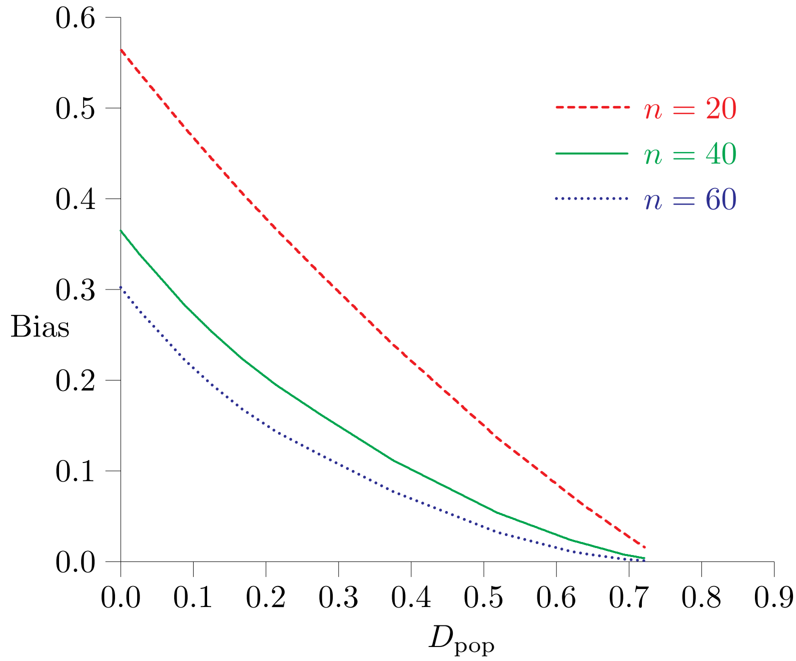

Figure 1: Bias E(D)-D_{pop}, $J = 4$, $p = 0.1$, equal expected unit sizes

3 Bias Correction

The purpose of a bias correction to D is to reduce the upward bias of the estimate of D_{pop}, as highlighted in Figure 1. We first consider a bootstrap bias correction, as described in Hall (1992) and Davison and Hinkley (1997) among many others. Given an observed allocation, a new sample is generated with the same sample size n and minority proportion p, but using the observed conditional probabilities $\hat{p}_j^1 = n_j^1/n^1$ and $\hat{p}_j^0 = n_j^0/n^0$ for the allocation process. Note that none of the bootstrap data-generating processes used in this paper involves resampling. The value for D in this bootstrap sample is denoted D_b. Repeating this B times, we can calculate

$$\bar{D}_b = \frac{1}{B}\sum_{b=1}^{B} D_b. \tag{3}$$

The population value of the segregation measure in the bootstrap sample is D itself, and so a measure of the bias of D is given by $\bar{D}_b - D$. A bootstrap bias-corrected estimate of D_{pop} is then obtained as

$$D_{bc} = D - (\bar{D}_b - D) = 2D - \bar{D}_b. \tag{4}$$

This type of bias correction works well if the bias is constant for different values of D_{pop}. This is clearly not the case here, as the biases as displayed in Figure 1 are much larger for smaller values of D. This bias correction is therefore not expected to work well for small unit sizes combined with small values of D_{pop}.

What turns out to be a more effective way of reducing the bias is a modified maximum-likelihood approach. As $n^c \to \infty$ with J fixed, $\hat{p}_j^c \to p_j^c$ almost surely, and the limiting distribution of $\sqrt{n^c}(\hat{p}_j^c - p_j^c)$ is normal with expectation 0 and variance $p_j^c(1-p_j^c)$. It follows that D is consistent for D_{pop}, and that it is asymptotically normal with an asymptotic variance that can be computed using the delta method, as will be pursued in Section 5. For bias reduction, rather than working with the full log-likelihood (2), we proceed as though the $\hat{p}_j^c$, $c = 1, 2$, $j = 1, \ldots, J$ were actually distributed according to their asymptotic normal distribution.

However, the asymptotic normality of D is not of the usual kind where the limiting distribution of $\sqrt{n}(D - D_{\text{pop}})$ would be normal with expectation zero. We have seen that $D = \sum_{j=1}^J |\hat{p}_j^1 - \hat{p}_j^0|/2$. The fact that D depends on the absolute values of the differences means that the expectation of the limiting distribution is not zero whenever the true value of D_{pop} is either zero or is such that $D_{\text{pop}} = O(n^{-1/2})$ as $n \to \infty$. This implies that asymptotic inference of the usual sort is not valid, since the non-zero expectation acts like a non-centrality parameter. But we can still use the asymptotic normal distribution of the $\hat{p}_j^c$ as an approximation in computing the asymptotic bias. The finite-sample bias is given by

$$\text{E}(D) - D_{\text{pop}} = \frac{1}{2} \sum_{j=1}^{J} \left(\text{E}|\hat{p}_j^1 - \hat{p}_j^0| - |p_j^1 - p_j^0| \right),$$

where, in each term of the sum, the random variables $\hat{p}_j^1$ and $\hat{p}_j^0$ are independent.

Let $X \sim \text{N}(\mu, \sigma^2)$. The distribution of $Z = |X|/\sigma$ is the so-called folded normal distribution, see Leone, Nelson and Nottingham (1961). Let θ denote $|\mu|/\sigma$. Then the density of Z is given by

$$f(z) = \phi(z - \theta) + \phi(z + \theta). \tag{5}$$

Here ϕ is the standard normal density. In order to derive the bias of D, we replace Z by $\hat{\theta}_j = |\hat{p}_j^1 - \hat{p}_j^0|/\hat{\sigma}_j$, where $\hat{p}_j^1 - \hat{p}_j^0 \sim \text{N}(p_j^1 - p_j^0, \sigma_j^2)$ approximately, with

$$\sigma_j^2 = p_j^1(1 - p_j^1)/n^1 + p_j^0(1 - p_j^0)/n^0 \quad \text{and} \quad \hat{\sigma}_j^2 = \hat{p}_j^1(1 - \hat{p}_j^1)/n^1 + \hat{p}_j^0(1 - \hat{p}_j^0)/n^0.$$

The asymptotic approximation to the density of $\hat{\theta}_j$ is then $\phi(\hat{\theta}_j - \theta_j) + \phi(\hat{\theta}_j + \theta_j)$, where $\theta_j = |p_j^1 - p^0|/\sigma_j$, the "true" value of θ. Thus, asymptotically, the data-dependent quantity $\hat{\theta}_j$ is sufficient for θ_j, and so an approximate maximum-likelihood estimate of θ_j is the value of θ_j that maximises the approximate density.

It can be shown that, for $\hat{\theta}_j \leq 1$, the maximum occurs at $\theta_j = 0$, and that the maximising θ_j tends to $\hat{\theta}_j$ as $\hat{\theta}_j \to \infty$. Let the maximising θ_j be denoted by $n(\hat{\theta}_j)$. The cut-off imposed by the function n is at $\hat{\theta}_j = 1$. Since $|\hat{p}_j^1 - \hat{p}_j^0| = \hat{\sigma}_j \hat{\theta}_j$, we can define another estimator of D_{pop}:

$$D_{dc} = \frac{1}{2} \sum_{j=1}^{J} \hat{\sigma}_j n(\hat{\theta}_j), \tag{6}$$

where the notation "*dc*" stands for "density-corrected".

We show in the next sections that the proposed bias correction procedures reduce enough of the bias to make reasonable inferences about levels of segregation, provided unit sizes are not too small. Where unit sizes are very small, we show in Section 4 that the observed level of segregation can rarely statistically be distinguished from evenness. Thus we suggest that in these cases the data are inappropriate for making inferences about segregation.

In the Appendix, we consider two other plausible methods of bias reduction, but simulations show that they are less effective than the density-correction method.

3.1 Monte Carlo Simulations

This section evaluates the performance of the bias adjustments for estimating levels of segregation. To do this we follow Duncan and Duncan's (1955) approach of generating a level of unevenness between no segregation and complete segregation using a single parameter, $0 \leq q < 1$. This parameter maps to a set of parabolic segregation curves via the formula:[2]

$$\Pr(\text{unit} \leq j \mid c = 1) = \frac{(1-q)\Pr(\text{unit} \leq j \mid c = 0)}{1 - q \cdot \Pr(\text{unit} \leq j \mid c = 0)}.$$

This formula, combined with the constraint of equal expected unit sizes, fixes the conditional allocation probabilities for both groups. An allocation is then generated by assigning n^1 and n^0 individuals to the J units using these calculated conditional probabilities.

This process is repeated 5,000 times for each n, p, and D_{pop} combinations over the following parameter space:

- Number of units, J, is fixed at 50;
- Unit sizes n_j are equal in expectation, with expected unit size 10, 30, or 50;
- Proportion of $c = 1$ individuals, p, equal to 0.05, 0.2 or 0.35;
- Systematic segregation generator, q, varies such that the values of D_{pop} are equal to 0, 0.056, 0.127, 0.225, 0.382 or 0.634.

For the bootstrap bias correction, $\bar{D}_b$ is calculated from (3) using 250 bootstrap samples. The bias and root mean squared error (RMSE) of D, D_{bc} from (4), and D_{dc} from (6) are presented in Table 1. It shows that, where the minority proportion is very small, $p = 0.05$, unit sizes are small (*e.g.* $\mathrm{E}(n_j) = 10$), and systematic segregation is very low (*e.g.* $D_{\text{pop}} = 0.056$), observed segregation incorrectly suggests that a highly segregating process underlies the allocation, with $D = 0.606$. The bias corrections, although reducing the bias, cannot fully get rid of it, the smallest bias being obtained with the density-corrected estimator, $D_{dc} = 0.406$. At the other extreme, where the minority proportion is large (*e.g.* $p = 0.35$), unit sizes are large (*e.g.* $n = 50$) and systematic segregation is high (*e.g.* $D_{\text{pop}} = 0.634$), no correction is needed, because the mean value of observed segregation is only slightly different from D_{pop}. However, in much social science data, the phenomenon of interest tends to have moderate (D_{pop} around 0.1 to 0.4) rather than very high levels of segregation. In this range, the bias corrections tend to work well and are necessary, provided that p and $\mathrm{E}(n_j)$ are not both simultaneously very small. For example, when the minority proportion is 10% and unit sizes are expected to be 30, if underlying segregation is 0.225, the observed index of segregation would be upward biased by 0.093 whereas the density-corrected estimator would successfully reduce this bias to just 0.005.

---

[2] Although this set of segregation curves cannot represent all distributions of segregation, it is a sufficient set to examine different levels of systematic segregation for the purposes of this paper.

The bias-corrected estimator D_{bc} is dominated in both bias and RMSE by the density-corrected estimator D_{dc} in almost all experiments, except for the cases of high D_{pop} values and larger minority proportions, in which the bias and RMSE of both corrected estimates are small.

Table 1: Bias and Root Mean Squared Error of D and bias-corrected estimators for $J = 50$ and combinations of p, $\mathrm{E}(n_j)$, and D_{pop}

$\mathrm{E}(n_j) = 10$	D_{pop}											
	0		0.056		0.127		0.225		0.382		0.634	
	Bias	*RMSE*	Bias	*RMSE*	Bias	*RMSE*	Bias	*RMSE*	Bias	*RMSE*	Bias	*RMSE*
$p = 0.05$												
D	0.60	*0.61*	0.55	*0.55*	0.48	*0.49*	0.40	*0.40*	0.29	*0.29*	0.15	*0.15*
D_{bc}	0.48	*0.49*	0.43	*0.43*	0.37	*0.37*	0.29	*0.30*	0.20	*0.20*	0.097	*0.11*
D_{dc}	0.40	*0.41*	0.35	*0.35*	0.29	*0.29*	0.21	*0.22*	0.13	*0.14*	0.058	*0.086*
$p = 0.10$												
D	0.41	*0.42*	0.36	*0.36*	0.30	*0.30*	0.23	*0.24*	0.15	*0.16*	0.071	*0.084*
D_{bc}	0.26	*0.27*	0.21	*0.22*	0.15	*0.17*	0.097	*0.12*	0.043	*0.077*	0.009	*0.058*
D_{dc}	0.26	*0.27*	0.21	*0.22*	0.15	*0.16*	0.094	*0.11*	0.040	*0.072*	0.011	*0.056*
$p = 0.20$												
D	0.31	*0.31*	0.26	*0.26*	0.20	*0.21*	0.15	*0.15*	0.089	*0.097*	0.039	*0.053*
D_{bc}	0.19	*0.20*	0.14	*0.15*	0.090	*0.11*	0.046	*0.067*	0.011	*0.051*	-0.002	*0.044*
D_{dc}	0.17	*0.18*	0.12	*0.13*	0.070	*0.082*	0.024	*0.052*	-0.009	*0.050*	-0.015	*0.047*
$p = 0.35$												
D	0.26	*0.26*	0.21	*0.21*	0.16	*0.16*	0.11	*0.11*	0.063	*0.072*	0.026	*0.042*
D_{bc}	0.16	*0.16*	0.11	*0.12*	0.063	*0.074*	0.027	*0.050*	0.004	*0.043*	-0.002	*0.038*
D_{dc}	0.15	*0.15*	0.095	*0.10*	0.048	*0.060*	0.009	*0.041*	-0.013	*0.045*	-0.012	*0.040*

Notes: Bias and Root mean squared error reported for 5000 replications. Number of bootstrap repetitions 250.

Table 1 continued

	D_{pop}											
$\mathrm{E}(n_j) = 30$	0		0.056		0.127		0.225		0.382		0.634	
	Bias	*RMSE*	Bias	*RMSE*	Bias	*RMSE*	Bias	*RMSE*	Bias	*RMSE*	Bias	*RMSE*
$p = 0.05$												
D	0.33	*0.34*	0.28	*0.28*	0.22	*0.23*	0.16	*0.17*	0.099	*0.11*	0.044	*0.057*
D_{bc}	0.21	*0.21*	0.16	*0.16*	0.10	*0.11*	0.055	*0.074*	0.015	*0.054*	-0.003	*0.046*
D_{dc}	0.18	*0.18*	0.13	*0.13*	0.072	*0.084*	0.024	*0.054*	-0.009	*0.053*	-0.015	*0.049*
$p = 0.10$												
D	0.24	*0.24*	0.19	*0.19*	0.14	*0.14*	0.093	*0.098*	0.052	*0.061*	0.022	*0.036*
D_{bc}	0.14	*0.15*	0.095	*0.10*	0.051	*0.063*	0.019	*0.043*	0.000	*0.040*	-0.003	*0.034*
D_{dc}	0.13	*0.14*	0.084	*0.089*	0.038	*0.051*	0.005	*0.038*	-0.010	*0.041*	-0.008	*0.035*
$p = 0.20$												
D	0.18	*0.18*	0.13	*0.13*	0.088	*0.090*	0.054	*0.059*	0.029	*0.039*	0.012	*0.026*
D_{bc}	0.11	*0.11*	0.060	*0.065*	0.024	*0.038*	0.005	*0.031*	-0.001	*0.031*	-0.001	*0.025*
D_{dc}	0.099	*0.10*	0.051	*0.056*	0.014	*0.030*	-0.006	*0.031*	-0.008	*0.032*	-0.004	*0.026*
$p = 0.35$												
D	0.15	*0.15*	0.10	*0.11*	0.065	*0.068*	0.038	*0.044*	0.020	*0.030*	0.008	*0.021*
D_{bc}	0.090	*0.092*	0.045	*0.050*	0.014	*0.029*	0.002	*0.027*	-0.001	*0.026*	-0.000	*0.021*
D_{dc}	0.083	*0.086*	0.038	*0.043*	0.005	*0.024*	-0.007	*0.027*	-0.006	*0.027*	-0.003	*0.021*
$\mathrm{E}(n_j) = 50$												
$p = 0.05$												
D	0.26	*0.26*	0.21	*0.21*	0.15	*0.16*	0.11	*0.11*	0.060	*0.069*	0.026	*0.040*
D_{bc}	0.15	*0.16*	0.11	*0.11*	0.061	*0.072*	0.024	*0.048*	0.003	*0.042*	-0.003	*0.035*
D_{dc}	0.15	*0.15*	0.098	*0.10*	0.052	*0.063*	0.016	*0.042*	-0.003	*0.041*	-0.005	*0.035*
$p = 0.10$												
D	0.19	*0.19*	0.14	*0.14*	0.093	*0.096*	0.058	*0.063*	0.031	*0.041*	0.013	*0.027*
D_{bc}	0.11	*0.11*	0.064	*0.070*	0.027	*0.040*	0.007	*0.032*	-0.001	*0.031*	-0.001	*0.026*
D_{dc}	0.10	*0.11*	0.056	*0.061*	0.017	*0.032*	-0.003	*0.031*	-0.007	*0.032*	-0.004	*0.027*
$p = 0.20$												
D	0.14	*0.14*	0.093	*0.094*	0.057	*0.060*	0.033	*0.038*	0.017	*0.027*	0.008	*0.020*
D_{bc}	0.082	*0.085*	0.039	*0.044*	0.011	*0.026*	0.001	*0.024*	-0.001	*0.023*	0.000	*0.020*
D_{dc}	0.076	*0.078*	0.032	*0.037*	0.003	*0.023*	-0.006	*0.025*	-0.005	*0.024*	-0.002	*0.020*
$p = 0.35$												
D	0.12	*0.12*	0.072	*0.073*	0.041	*0.044*	0.023	*0.029*	0.012	*0.022*	0.005	*0.016*
D_{bc}	0.069	*0.071*	0.027	*0.032*	0.005	*0.021*	-0.000	*0.020*	-0.000	*0.020*	-0.000	*0.017*
D_{dc}	0.064	*0.066*	0.021	*0.027*	-0.002	*0.020*	-0.006	*0.022*	-0.003	*0.020*	-0.001	*0.017*

Notes: Bias and Root mean squared error reported for 5000 replications. Number of bootstrap repetitions 250.

4 Tests of No Systematic Segregation

To complement the bias-corrected estimators of D, we provide a test for no systematic segregation. We consider two alternative methods to test whether we can reject the hypothesis that the level of segregation observed was generated by randomness alone, $D_{\text{pop}} = 0$. It is common in the literature to run a randomisation procedure to generate the distribution of D under the null of no systematic segregation (see *e.g.* Boisso *et al.*, (1994)), and D is compared to this distribution. Here, we generate the distribution of D under the null of no systematic segregation by creating B samples generated using the restricted conditional probabilities $\hat{p}_j^0 = \hat{p}_j^1 = \hat{p}_j = (n_j^0 + n_j^1)/n$ and calculating D in each sample, which we denote D^*. The null hypothesis $H_0 : D_{\text{pop}} = 0$ is then rejected at level α if $1/B \sum_{b=1}^{B} \mathrm{I}(D_b^* > D) < \alpha$, where $\mathrm{I}(.)$ is the indicator function.
Alternatively, following the statistical model developed in Section 2, we can employ a likelihood ratio test for the hypothesis

$$H_0 : p_j^0 = p_j^1 = p_j \quad \forall j,$$

with test statistic

$$LR = 2\sum_{j=1}^{J} \left[n_j^0 \log \hat{p}_j^0 + n_j^1 \log \hat{p}_j^1 - n_j \log \hat{p}_j \right],$$

which follows an asymptotic χ^2_{J-1} distribution. This asymptotic distribution is for large n and fixed J, and therefore for large unit sizes. For large J and/or small unit sizes, the asymptotic approximation can be expected to be poor, as we originally found in our simulation results discussed below. We therefore also utilise a bootstrap procedure to improve the size properties of the test. Let LR^* be the value of the likelihood ratio test in a sample generated from $\hat{p}_j^0 = \hat{p}_j^1 = \hat{p}_j = (n_j^0 + n_j^1)/n$. Then the null hypothesis of no systematic segregation is rejected at level α if $1/B \sum_{b=1}^{B} \mathrm{I}(LR_b^* > LR) < \alpha$.
Table 2 presents the test results for $J = 50$ and $\mathrm{E}(n_j) = 10$ and $\mathrm{E}(n_j) = 30$, for various values of D_{pop} and minority proportions p. The number of Monte Carlo replications was 10,000 with 599 bootstrap samples. The size and power properties of the two tests are virtually identical. They have good size properties for all minority proportions p, with overall LR dominating the Randomisation test. The tests fail to reject the null for small values of D_{pop} combined with small minority proportions p, exactly the circumstances in which the bias corrections do not remove much of the bias of D. Clearly, any calculation of D and the bias-corrected estimators should be accompanied by the D^* and/or bootstrapped LR tests. If these tests fail to reject, no further inference should be pursued.

5 Inference on D

Having established that the bias corrections work well for a large part of the parameter space, we next develop reliable inference procedures such as 95% confidence intervals and Wald test statistics for equivalence of segregation in different areas. Inference

Table 2: Rejection frequencies of D Randomisation and Likelihood Ratio tests, for $J = 50$, level $\alpha = 0.05$.

$\mathrm{E}(n_j) = 10$							
p	Test	D_{pop}					
		0	**0.056**	**0.127**	**0.225**	**0.382**	**0.634**
0.05	D^*	0.096	0.104	0.131	0.237	0.619	0.998
	LR	0.066	0.074	0.098	0.194	0.594	0.999
0.10	D^*	0.056	0.069	0.112	0.307	0.878	1.000
	LR	0.069	0.083	0.132	0.362	0.919	1.000
0.20	D^*	0.067	0.086	0.192	0.618	0.999	1.000
	LR	0.062	0.080	0.183	0.606	0.998	1.000
0.35	D^*	0.065	0.090	0.269	0.827	1.000	1.000
	LR	0.053	0.077	0.232	0.791	1.000	1.000
$\mathrm{E}(n_j) = 30$							
0.05	D^*	0.060	0.071	0.165	0.534	0.992	1.000
	LR	0.051	0.067	0.160	0.546	0.995	1.000
0.10	D^*	0.056	0.086	0.285	0.882	1.000	1.000
	LR	0.054	0.080	0.275	0.877	1.000	1.000
0.20	D^*	0.057	0.117	0.553	0.997	1.000	1.000
	LR	0.050	0.108	0.537	0.997	1.000	1.000
0.35	D^*	0.055	0.147	0.775	1.000	1.000	1.000
	LR	0.050	0.138	0.777	1.000	1.000	1.000

Notes: Rejection frequencies reported for 10,000 replications. Number of bootstrap repetitions 599.

based on a bias corrected estimator is of course expected to work well only in that part of the parameter space where the bias corrections work well, i.e. where the tests of no systematic segregation reject the null strongly, as indicated in Table 2.

We start by deriving the asymptotic distribution of D given our statistical framework, following the procedures as developed in Ransom (2000).
Under the data generating process as described in Section 2, for $0 < p_j^c < 1$, with $c = 0, 1$; $j = 1, ..., J$; $\sum_j p_j^c = 1$, the estimated conditional probabilities $\hat{p}_j^c$, are asymptotically normally distributed, as

$$\sqrt{n^c}\begin{bmatrix}\hat{p}_1^c - p_1^c \\ \hat{p}_2^c - p_2^c \\ \vdots \\ \hat{p}_J^c - p_J^c\end{bmatrix} = \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} p_1^c(1-p_1^c) & -p_1^c p_2^c & \dots & -p_1^c p_J^c \\ -p_1^c p_2^c & p_2^c(1-p_2^c) & \dots & -p_2^c p_J^c \\ \vdots & \vdots & \ddots & \vdots \\ -p_1^c p_J^c & -p_2^c p_J^c & \dots & p_J^c(1-p_J^c)\end{bmatrix}\right) \equiv \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega}^c).$$

As $n^1 = pn$ and $n^0 = (1-p)n$, the limiting distribution of D can then be obtained via the delta method:

$$\sqrt{n}(D - D_{\mathrm{pop}}) \xrightarrow{d} \mathrm{N}\big(0, \lambda^\top(p^{-1}\boldsymbol{\Omega}^1 + (1-p)^{-1}\boldsymbol{\Omega}^0)\lambda\big),$$

where λ is a J-vector with r^{th} element $\lambda_r = \mathrm{sign}(p_r^1 - p_r^0)/2$, where $\mathrm{sign}(q) = 1$ if $q > 0$ and $\mathrm{sign}(q) = -1$ of $q < 0$. [3] This follows from

$$\frac{\partial D_{\mathrm{pop}}}{\partial p_r^1} = \frac{\partial}{\partial p_r^1}\frac{1}{2}\sum_{j=1}^{J}|p_j^1 - p_j^0| = \mathrm{sign}(p_r^1 - p_r^0)/2, \text{ and}$$

$$\frac{\partial D_{\mathrm{pop}}}{\partial p_r^0} = \frac{\partial}{\partial p_r^0}\frac{1}{2}\sum_{j=1}^{J}|p_j^1 - p_j^0| = -\,\mathrm{sign}(p_r^1 - p_r^0)/2.$$

Clearly, this derivation is valid only when $|p_r^1 - p_r^0|$ is not in a root-n neighbourhood of zero, as discussed in Section 3. The asymptotic distribution of D is then given by

$$D \overset{a}{\sim} \mathrm{N}\big(D_{\mathrm{pop}}, n^{-1}\lambda^\top(p^{-1}\boldsymbol{\Omega}^1 + (1-p)^{-1}\boldsymbol{\Omega}^0)\lambda\big),$$

or, equivalently,

$$D \overset{a}{\sim} \mathrm{N}\big(D_{\mathrm{pop}}, \lambda^\top(\boldsymbol{\Omega}^1/n^1 + \boldsymbol{\Omega}^0/n^0)\lambda\big),$$

which can form the basis for constructing confidence intervals and Wald test statistics for hypotheses of the form $H_0 : D_{\mathrm{pop}} = \delta$. If we denote by $\hat{\lambda}$ and $\hat{\boldsymbol{\Omega}}^c$ the estimated counterparts of λ and $\boldsymbol{\Omega}^c$, and substitute the observed fractions $\hat{p}_j^c$ for p_j^c, the Wald test is then computed as

$$W = \frac{(D-\delta)^2}{\hat{\lambda}^\top(\hat{\boldsymbol{\Omega}}^1/n^1 + \hat{\boldsymbol{\Omega}}^0/n^0)\hat{\lambda}}, \tag{7}$$

[3] Although $\boldsymbol{\Omega}^c$ is singular because $\sum_j p_j^c = 1$, exactly the same results are obtained by redefining D as a function of $2(J-1)$ probabilities only.

and converges in distribution to the χ^2_1 distribution under the null.

Clearly, we don't expect this approximation to work well when δ, group sizes, and/or minority proportions are small, if only on account of the upward bias of D as established in the previous sections. However, the Wald test statistic W is asymptotically pivotal in the sense that its limiting distribution is not a function of nuisance parameters. We can therefore use bootstrap P values, which may result in an improvement in the finite-sample behaviour of the test; see for instance Hall (1992), Davison and Hinkley (1997), Davidson and MacKinnon (2000) and Davidson (2009). If we denote the Wald statistic in the b^{th} bootstrap sample as W_b, calculated as

$$W_b = \frac{(D_b - D)^2}{\hat{\lambda}_b{}^\top(\hat{\boldsymbol{\Omega}}^1_b/n^1 + \hat{\boldsymbol{\Omega}}^0_b/n^0)\hat{\lambda}_b}, \tag{8}$$

the bootstrap P value is then given by $1/B\sum_{b=1}^{B}\mathrm{I}(W_b > W)$. This bootstrap procedure is equivalent to a symmetric two-tailed test for the t statistic.

Let τ denote the t statistic that is the signed square root of the Wald statistic (7). Let τ_b denote the signed square root of (8). Then a test that does not assume symmetry can be based on the equal-tail bootstrap P value

$$2\min\Big[\frac{1}{B}\sum_{b=1}^{B}\mathrm{I}(\tau_b < \tau), \frac{1}{B}\sum_{b=1}^{B}\mathrm{I}(\tau_b > \tau)\Big].$$

Alternatively, we can base the inference directly on any of the bias-corrected estimators of D_{pop}. In order to estimate the variance of the bias-corrected estimators, we again perform a bootstrap procedure. For example, denoting the bootstrap estimate of the variance of D_{dc} by $\widehat{\mathrm{Var}}_b(D_{dc})$, the Wald test based on D_{dc} is then calculated as

$$W_{dc} = \frac{(D_{dc} - \delta)^2}{\widehat{\mathrm{Var}}_b(D_{dc})},$$

and this is again compared to the χ^2_1 distribution.

Figure 2 shows P value plots for testing the true hypothesis $H_0 : D_{\text{pop}} = 0.292$, for $\mathrm{E}(n_j) = 30$, $J = 50$, and $p = 0.3$. The Wald test that is based on the asymptotic normal distribution of D and uses χ^2_1 critical values is denoted W, whereas the Wald test that uses the bootstrap critical values is denoted W_{pb}. The test based on the equal-tail bootstrap P value for the t test is denoted T_{pb}. The Wald statistic that uses the density-corrected estimator and its bootstrap variance estimate is denoted W_{dc}. The results shown are for 10,000 Monte Carlo replications. 599 bootstrap samples per replication are drawn for the calculation of the variances and the bootstrap distribution of the Wald test.

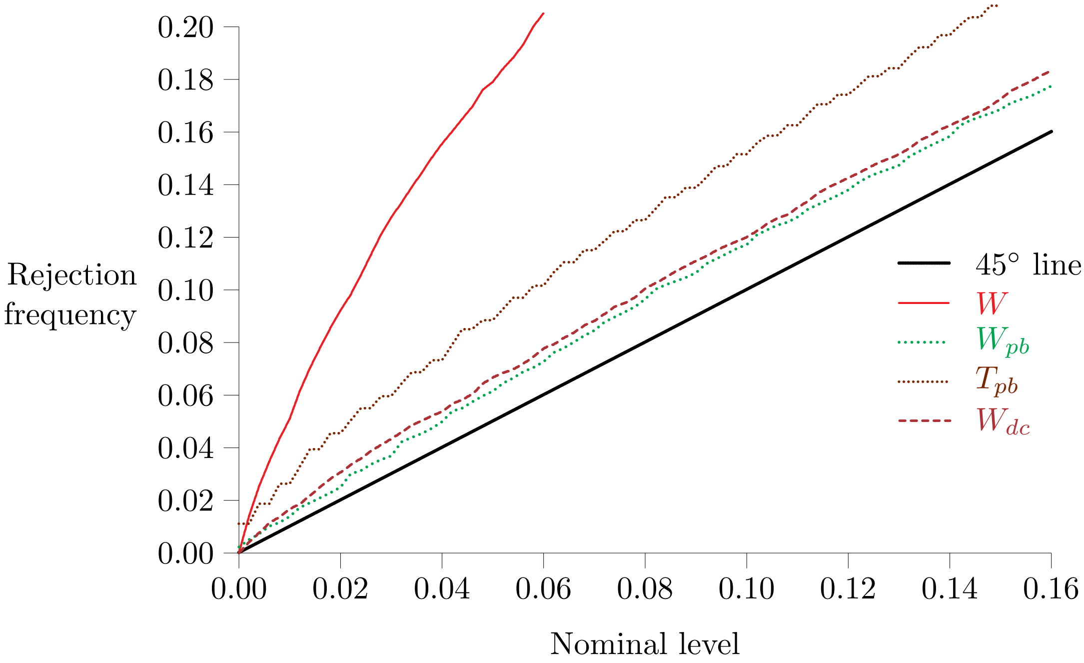

Figure 2: P value plot, H_0: $D_{\text{pop}} = 0.292$, $\text{E}(n_j) = 30$, $J = 50$, $p = 0.30$.

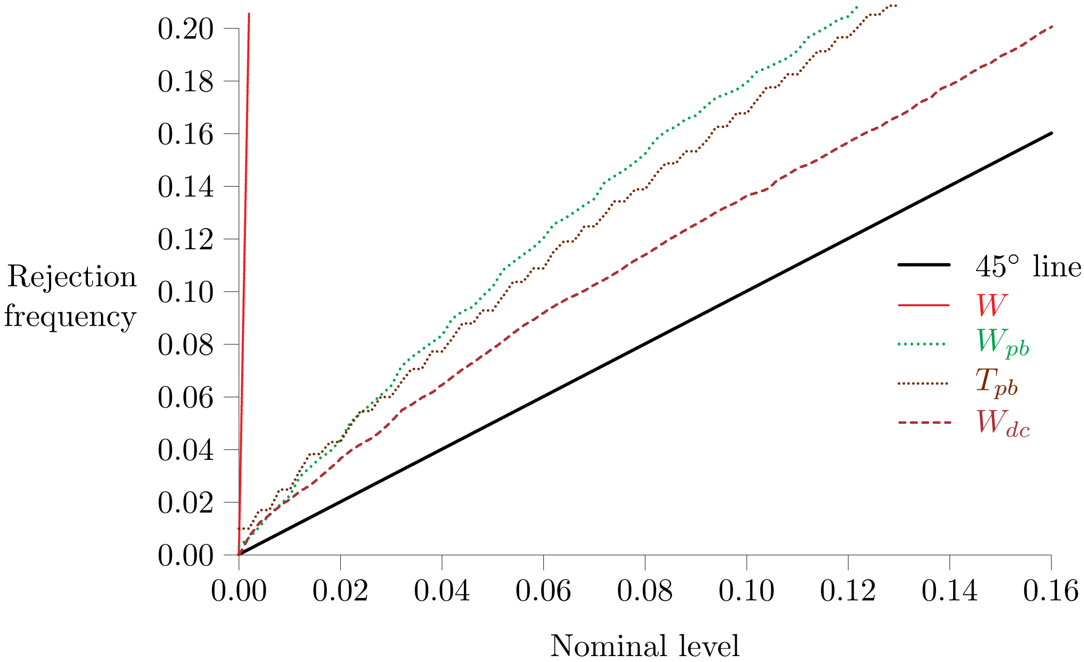

Figure 3: P value plot, H_0: $D_{\text{pop}} = 0.292$, $\text{E}(n_j) = 20$, $J = 50$, $p = 0.10$.

The first column of numbers in Table 3 presents the bias and root mean squared error for the various estimates. There is a 11% upward bias in D, but D_{bc} is unbiased. D_{dc} has a small downward bias of 2.8%. As is clear from Figure 2, the asymptotic Wald test based on W using the χ^2_1 critical values does not have good size properties. It rejects the true null too often, for example at 5% nominal size, it rejects the null in 18.6% of the replications. In contrast, using the P values from the bootstrap distribution of the Wald statistic improves the size behaviour considerably. At the 5% level, the rejection frequency is now reduced to 6.9%. Using the equal-tailed bootstrap P values for the t test also improves on the size performance of the asymptotic Wald statistic, but it performs less well than W_{pb}. W_{dc} has the same size properties as W_{pb}.

Table 3: Bias and RMSE of D and bias-corrected estimators

	$D_{\text{pop}} = 0.292$, $\text{E}(n_j) = 30$, $p = 0.30$		$D_{\text{pop}} = 0.292$, $\text{E}(n_j) = 20$, $p = 0.10$		$D_{\text{pop}} = 0.382$, $\text{E}(n_j) = 30$, $p = 0.30$	
	Bias	*RMSE*	Bias	*RMSE*	Bias	*RMSE*
D	0.031	*0.038*	0.106	*0.111*	0.022	*0.032*
D_{bc}	-0.000	*0.027*	0.020	*0.051*	-0.001	*0.026*
D_{dc}	-0.008	*0.029*	-0.000	*0.046*	-0.007	*0.028*

Notes: $J = 50$ in all designs. Results from 10,000 Monte Carlo replications.

Figure 3 shows the P value plot for a similar design, but now for smaller expected group sizes $\text{E}(n_j) = 20$ and a smaller minority proportion, $p = 0.10$. The bias of D in this case is 0.106, or 36%, that of D_{bc} is around 0.020, or 6.5%, while D_{dc} is unbiased.
The size distortions of the test statistics are now more severe. The asymptotic Wald test is severely size distorted, with a 68% rejection rate at the 5% level. The Wald and asymmetric t test using the bootstrap P values behave much better, with T_{pb} behaving somewhat better. At the 5% level, the rejection frequencies for these tests are 10% and 9.0% respectively. W_{dc} here has the best size performance of all tests; it rejects the true null 7.0% of the time at the 5% level and is the only test where the size properties remain the same as those seen in Figure 2.
There is a one-to-one correspondence between the P value plots as depicted in Figures 2 and 3 and the coverage properties of the confidence intervals associated with the particular test statistics. Using the normal approximation, $(1-\alpha)\%$ confidence intervals associated with the asymptotic Wald and W_{bc} tests are constructed as

$$D - z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(D)} < D_{\text{pop}} < D + z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(D)}$$

and

$$D_{bc} - z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}_b(D_{bc})} < D_{\text{pop}} < D_{bc} + z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}_b(D_{bc})}$$

respectively, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. For the bootstrap Wald test the associated confidence interval is given by

$$D - \sqrt{w^*_{1-\alpha}\widehat{\text{Var}}(D)} < D_{\text{pop}} < D + \sqrt{w^*_{1-\alpha}\widehat{\text{Var}}(D)},$$

where $w^*_{1-\alpha}$ is the $1-\alpha$ quantile of the distribution of the bootstrap repetitions W_b. The equal-tailed bootstrap t test has the corresponding confidence interval given by

$$D - \tau^*_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(D)} < D_{\text{pop}} < D - \tau^*_{\alpha/2}\sqrt{\widehat{\text{Var}}(D)},$$

where $\tau^*_{1-\alpha/2}$ and $\tau^*_{\alpha/2}$ are respectively the $1-\alpha/2$ and $\alpha/2$ quantiles of the distribution of the bootstrap repetitions τ_b.
For the example with $\text{E}(n_j) = 20$ and $p = 0.10$ as described above, the observed rejection frequencies of 68%, 9.8%, 9.0% and 7.0% for the W, W_{pb}, T_{pb}, and W_{dc} tests respectively translate into coverage probabilities of 32%, 90.2%, 91% and 93% of the associated 95% confidence intervals. Given the upward bias of D this leads to an interesting observation concerning the confidence interval based on the bootstrap Wald test W_{pb}. As the size and associated coverage properties of this test are reasonably good, but as the confidence interval is symmetric around the upward biased D, this suggests that the W_{pb} based confidence interval will be quite large. Table 4 shows the averages of the lower and upper limits and lengths of the 95% confidence intervals based associated with W_{pb}, T_{pb}, and W_{dc} respectively. This confirms that the W_{pb}-based confidence interval is on average indeed much wider than those based on W_{dc} and T_{pb}. Whereas the lower limit is quite similar to those of the other two confidence intervals, its upper limit is much higher, as expected owing to the symmetry around the upward biased D. Clearly, W_{pb} can therefore have poor power properties when D has substantial bias.
In principle, a likelihood ratio test of the hypothesis $D_{\text{pop}} = \delta$ is possible. In practice, obtaining the maximised log-likelihood under that constraint is a largely intractable problem, since the constraint is not differentiable with respect to p^0_j and p^1_j when they are equal.

Table 4: Average lower limit, upper limit and length of 95% confidence intervals

Test	Lower limit	Upper limit	Length
W_{pb}	0.228	0.568	0.340
T_{pb}	0.212	0.378	0.166
W_{dc}	0.209	0.376	0.167

Notes: $D_{\text{pop}} = 0.292$, $\text{E}(n_j) = 20$, $J = 50$, $p = 0.10$.

The test results presented here show that inference can be based on the W_{pb}, T_{pb}, and W_{dc} tests when the sample size, the value of D_{pop}, and the minority proportion are such that the bias corrections work reasonably well, although as Figures 2 and 3 show, some size distortions occur also for these tests. We next consider the case where $D_{\text{pop}} = 0.127$, $E(n_j) = 10$ and $p = 0.10$. From Table 1 it is clear that the bias corrected estimators remain biased in this case, with D_{dc} having a bias of 0.15. Table 2 also shows that the tests for no systematic segregation only reject the null around 12% of the time. The rejection frequencies at the 5% level for the W, W_{pb}, T_{pb}, and W_{dc} tests are 100%, 42%, 26% and 80% respectively, indicating that W_{dc} in this case is severely oversized, as expected, whereas T_{pb} performs best, but is still oversized substantially. When we increase $\text{E}(n_j)$ to 30, the bias of D_{dc} is reduced to 0.038, and the tests reject the null of no systematic segregation around 28% of the time. The rejection frequencies at the 5% level for W, W_{pb}, T_{pb}, and W_{dc} are now 99%, 35%, 25% and 16% respectively, showing a substantial improvement of the performance of W_{dc}, whereas the performance of T_{pb} is similar to before. At the other end of the scale, when we set $D_{\text{pop}} = 0.634$, $E(n_j) = 30$ and $p = 0.35$, all tests work well with rejection frequencies at the 5% level of 6.6%, 5.2%, 6.5% and 5.2% for W, W_{pb}, T_{pb}, and W_{dc} respectively.

6 Tests for Equality of Segregation

A researcher may well be interested in determining whether segregation has changed significantly within an area over time, or whether segregation in one area is significantly different from that in another, similar, or perhaps neighbouring area. We consider the performances of the test statistics for comparing the two hypothetical areas for which the results were simulated above. Area 1 has $J = 50$, $\text{E}(n_j) = 30$ and $p = 0.30$, whereas Area 2 has $J = 50$, $\text{E}(n_j) = 20$ and $p = 0.10$. To study the size properties of the tests for the null hypothesis

$$H_0 : D_{\text{pop},1} = D_{\text{pop},2}$$

we set the two area population segregation measures $D_{\text{pop},1} = D_{\text{pop},2} = 0.292$ as before. Given the area-specific conditional allocation probabilities, the allocations in the areas are determined independently and therefore the Wald test

$$W = \frac{(D_1 - D_2)^2}{\widehat{\text{Var}}(D_1) + \widehat{\text{Var}}(D_2)}$$

is asymptotically χ^2_1 distributed. The Wald test based on *e.g.* the density-corrected estimates is defined as

$$W_{bc} = \frac{(D_{dc,1} - D_{dc,2})^2}{\widehat{\text{Var}}_b(D_{dc,1}) + \widehat{\text{Var}}_b(D_{dc,2})},$$

whereas the bootstrap P values for the W_{pb} test are based on the distribution of the bootstrap repetitions of

$$W_b = \frac{\big(D_{b,1} - D_{b,2} - (D_1 - D_2)\big)^2}{\widehat{\text{Var}}(D_{b,1}) + \widehat{\text{Var}}(D_{b,2})},$$

where $D_{b,1}$ and $D_{b,2}$ are calculated from independent bootstrap repetitions. The bootstrap P values for the T_{pb} test are obtained in an equivalent way.

As an example, the bias of $D_1 - D_2$ as an estimate for $D_{\text{pop},1} - D_{\text{pop},2}$ can be obtained from the results as presented in Table 3 and is equal to -0.075. As the covariance between D_1 and D_2 is equal to zero, the RMSE can be calculated as $((RMSE(D_1))^2 + (RMSE(D_2))^2 - 2Bias(D_1)Bias(D_2))^{1/2} = 0.085$. The equivalent numbers for $D_{dc,1} - D_{dc,2}$ are -0.007 and 0.054 for the bias and RMSE respectively. Note that this RMSE calculation is not exact, as the Monte Carlo sample covariance between the two area segregation measures is not exactly equal to zero, but this difference is negligible.

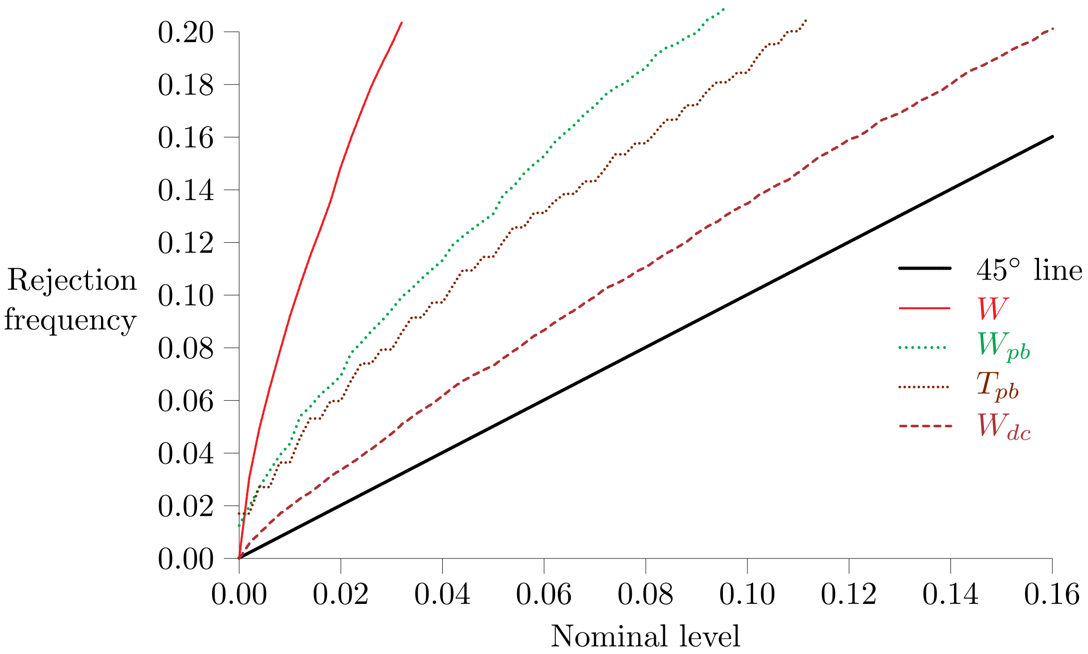

Figure 4: P value plot, $H_0 : D_{\text{pop},1} = D_{\text{pop},2}$, size properties

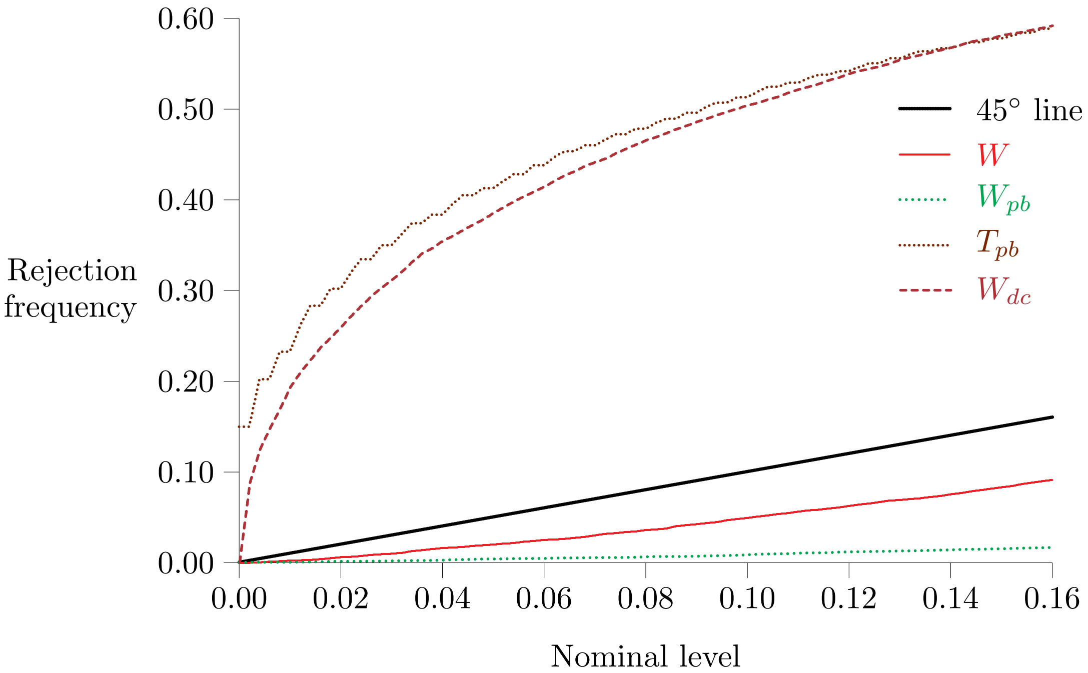

Figure 5: P value plot, $H_0 : D_{\text{pop},1} = D_{\text{pop},2}$, power properties

Figure 4 depicts the P value plots for the true null of equal population segregation measures D_{pop} in the two areas. The asymptotic Wald test again over-rejects substantially, 28.2% at the 5% level. The W_{dc} test displays the best size properties in this case, rejecting 8.6% of the time at the 5% level, followed by T_{pb}, and then W_{pb}. The rejection probabilities for these latter tests at the 5% level are 11.3% and 13.4% respectively.

We next turn to the power properties of these tests when the two population segregation measures $D_{\text{pop},1}$ and $D_{\text{pop},2}$ are not equal. We keep $D_{\text{pop},2}$ equal to 0.292, but increase $D_{\text{pop},1}$ to 0.382. The estimation results for this design are presented in the third column in Table 3. As discussed above, because D_2 is substantially biased upwards, we expect the W_{pb} test to have low power. This is confirmed by the P value plots in Figure 5. The standard Wald test has power below nominal size, but the bootstrap based Wald test W_{pb} completely fails to reject the null of equal segregation. In contrast, the Wald tests based on the bias-corrected estimator and T_{pb} show reasonable power properties, with T_{pb} having most power to detect this deviation from the null, although it has not been size adjusted. The P value plots, not shown here, for the true null that $D_{\text{pop},1} - D_{\text{pop},2} = 0.0897$ are very similar to those in Figure 4. Clearly, these results combined show that for simple hypothesis testing, W_{dc} and T_{pb} are the test procedures with reasonably good size and power properties in the settings we considered.

7 Discussion

Recently, Rathelot (2012) and d'Haultfœuille and Rathelot (2011) have also considered the problem of measuring segregation when units are small. In their setup, the number of individuals in units is small at around 5 or 10, whereas the number of units is large. Indeed the methods in these papers rely on large-number-of-units asymptotics. Both papers show that the parametric method proposed by Rathelot (2012) performs well in the estimation of the dissimilarity index and other measures of inequality, like the Gini and the Theil indices. In this setup, the number of individuals in unit j, n_j is drawn from a given, unknown distribution. Then the number of individuals in unit j having status c, n_j^c, is distributed as $\text{Binomial}\left(n_j, \pi_j^c\right)$, and $\widehat{\pi}_j^c = n_j^c/n_j$ is an unbiased estimate of π_j^c. The parametric method of Rathelot (2012) is then to assume that π_j^c is distributed as a mixture of Beta distributions, leading to the Beta-Binomial model. In the simulations in Rathelot (2012) it is shown that estimates of the Dissimilarity Index from this Beta-Binomial model have better properties in terms of smaller bias than the bootstrap bias correction in the setting of a large number of small units. We have tried to analyse the behaviour of this estimator in our setup, but found in the simulations of our design of Section 3.1 that the Beta-Binomial estimation procedure often did not converge, [4] making it difficult to compare the performance of this estimator with the others in our setup. For larger minority fractions and values of D_{pop} convergence of the ML estimator is easier to obtain in our design. Table 5 shows the estimation results for D_{beta1}, which is the Beta-Binomial estimator with a one-component Beta distribution, for $\text{E}(n_j) = 30$, $J = 50$, $p = 0.20$, where we obtained valid Monte

[4] We found this both using the R-programme as kindly provided to us by Roland Rathelot, as when using our own Gauss code.

Carlo results for $D_{\text{pop}} = 0.127$ and larger. The results show that in this design, D_{beta1} behaves similarly to D_{dc} for $D_{\text{pop}} = 0.127$, but has a substantially larger bias and/or RMSE for larger values of D_{pop}. For this design, increasing the number of mixtures of beta distributions to two does not change the results, as in almost all cases this model converges to the one-component model. The Beta-Binomial estimator is not consistent for a fixed number of units and unit sizes going to infinity. For example, when we increase the expected sample size to $\mathrm{E}(n_j) = 100$, the bias and RMSE of D_{beta1} increases to -0.042 and 0.044 respectively when $D_{\text{pop}} = 0.382$. The bias and RMSE of D itself in that case are 0.009 and 0.018 respectively.

Another bias corrected measure is the one proposed by Carrington and Troske (1997), which has been widely used in school segregation (Söderström and Uusitalo, 2005) and occupational segregation research (Hellerstein and Neumark, 2008). Carrington and Troske (1997) argue that segregation indices can be modified to take into account the underlying value under no systematic segregation, when $p_j^1 = p_j^0$, $j = 1, ..., J$. They propose a modified segregation index that measures the (economic) extent to which a sample deviates from the expected value of D under no systematic segregation, denoted $D^* = E\left(D\right)_{p_j^1 = p_j^0}$, which can be calculated as in Section 4. They argue that their new index of systematic dissimilarity does not depend on the margins in the area and is therefore a better means for comparing the extent to which systematic segregation exists. Their measure, denoted D_{ct} here, is defined as

$$D_{\text{ct}} = \frac{D - D^*}{1 - D^*} \text{ if } D \geq D^*; \qquad D_{\text{ct}} = \frac{D - D^*}{D^*} \text{ if } D < D^*,$$

and hence $D_{\text{ct}} \in [-1.1]$. D_{ct} can be interpreted as the extent to which the area is more dissimilar than random allocation would imply, expressed as a fraction of the maximum amount of such excess dissimilarity that could possibly occur. $D_{\text{ct}} = 0$ implies that the allocation of individuals in the area is equivalent to no systematic segregation. It is worth noting that D_{ct} is almost identical to the index proposed by Winship (1977), criticised by Falk et al. (1978), and partially withdrawn by Winship (1978). The problem with D_{ct} is that it is not entirely clear what it is intended to achieve. D_{ct} is always lower than D_{pop}, but tends to D_{pop} for large unit sizes. In our simulations, for many values of the parameters, D_{ct} underestimates D_{pop} by a larger amount than D itself overestimates D_{pop}. As an illustration of this, Table 5 presents simulation results for $\mathrm{E}(n_j) = 30$, $J = 50$, $p = 0.20$. The problem with D_{ct} is that decomposition of a segregation index into one part produced by systematic segregation and another part produced by randomness cannot be done additively.

In the previous Monte Carlo designs, all unit sizes were equal in expectation. We next perform the simulations with designs like those of the first two columns in Table 3, but for unequal expected unit sizes. In the first design, labelled Design 1 in Table 6, we split the 50 units into half with a smaller expected group size and half with a larger expected group size. For the 1500 individuals example, these sizes are approximately 10 and 50, for the second example of 1000 individuals, these are approximately 7 and 33. In the second design, labelled Design 2 in the table, all expected unit sizes are different, within the aforementioned ranges in steps of just under 1 individual. Table 6 and Table 7 present the estimation and testing results. The results are all very similar to those with equal expected unit sizes as presented in Sections 5 and 6.

Table 5: Bias and Root Mean Squared Error of D_{beta1} and D_{ct}

	D_{pop}											
	0		0.056		0.127		0.225		0.382		0.634	
	Bias	*RMSE*	Bias	*RMSE*	Bias	*RMSE*	Bias	*RMSE*	Bias	*RMSE*	Bias	*RMSE*
D	0.18	*0.18*	0.13	*0.13*	0.088	*0.090*	0.054	*0.059*	0.029	*0.039*	0.012	*0.026*
D_{dc}	0.099	*0.10*	0.051	*0.056*	0.014	*0.030*	-0.006	*0.031*	-0.008	*0.032*	-0.004	*0.026*
D_{beta1}					-0.012	*0.037*	-0.020	*0.037*	-0.033	*0.042*	-0.054	*0.060*
D_{ct}	-0.032	*0.075*	-0.067	*0.089*	-0.085	*0.091*	-0.10	*0.11*	-0.099	*0.10*	-0.064	*0.085*

Notes: $\text{E}(n_j) = 30$, $J = 50$, $p = 0.20$. No results reported for D_{beta1} in first two columns owing to convergence problems of the estimator.

Table 6: Results for designs with unequal expected unit sizes

	$n = 1500$, Bias	$p = 0.3$ *RMSE*	$n = 1000$, Bias	$p = 0.1$ *RMSE*
Design 1				
D	0.022	*0.032*	0.083	*0.091*
D_{bc}	-0.003	*0.027*	0.018	*0.050*
D_{dc}	-0.010	*0.029*	0.009	*0.047*
Design 2				
D	0.027	*0.035*	0.094	*0.100*
D_{bc}	-0.002	*0.027*	0.016	*0.049*
D_{dc}	-0.009	*0.028*	0.005	*0.045*
Rejection frequencies for tests of $H_0 : D_{\text{pop}} = 0.292$				
Nominal size	0.10	*0.05*	0.10	*0.05*
Design 1				
W	0.192	*0.116*	0.596	*0.457*
W_{pb}	0.106	*0.056*	0.200	*0.119*
T_{pb}	0.162	*0.100*	0.175	*0.100*
W_{dc}	0.135	*0.078*	0.127	*0.071*
Design 2				
W	0.246	*0.154*	0.726	*0.596*
W_{pb}	0.115	*0.064*	0.167	*0.096*
T_{pb}	0.161	*0.095*	0.167	*0.095*
W_{dc}	0.134	*0.075*	0.134	*0.076*

Notes: $J = 50$, $D_{\text{pop}} = 0.292$; 10,000 Monte Carlo replications, 599 bootstrap repetitions

Table 7: Test results for $H_0 : D_{\text{pop},1} = D_{\text{pop},2}$ sizes

	Design 1				Design 2			
Size	W	W_{pb}	T_{pb}	W_{dc}	W	W_{pb}	T_{pb}	W_{dc}
0.10	0.293	0.221	0.193	0.139	0.352	0.208	0.186	0.138
0.05	0.184	0.140	0.119	0.079	0.234	0.125	0.118	0.081

Notes: see Table 6

8 Social Segregation in Schools

In this section we illustrate our inference procedures with an empirical application relating to social segregation in primary schools in England. The dichotomous measure is an indicator of poverty based on eligibility for free school meals (FSM). This context is useful as it naturally produces small unit sizes, and shows a range of minority proportions and overall populations across different Local Authorities (LAs). We use administrative data collected by the Department for Children, Families and Schools and made available to researchers as part of the National Pupil Database on pupils aged 10/11 in English primary schools in 2006. Measurement of school segregation using this dataset has been carried out by many researchers; see Allen and Vignoles (2007), Burgess *et al.* (2006), and Gibbons and Telhaj (2006). Using the tools developed above, we can assess whether the small unit sizes and/or small minority populations lead to incorrect inferences about differences

Table 8: Key parameters of primary schools across English local authorities

LA name	Number of pupils	Number of schools	Average cohort size	% FSM	D
North-East Lincolnshire	2005	46	44	21	0.43
North Lincolnshire	2011	57	35	13	0.36
Blackburn	2105	51	41	26	0.34
Oldham	2990	86	35	21	0.47
Camden	1394	41	34	42	0.23
Greenwich	2666	66	40	36	0.29
Hackney	2194	54	41	43	0.22
Hammersmith & Fulham	1177	39	30	45	0.30
Islington	1845	48	38	41	0.26
Kensington & Chelsea	881	27	33	36	0.32
Lambeth	2428	60	40	40	0.24
Lewisham	2833	70	40	29	0.30
Southwark	2929	72	41	36	0.21
Tower Hamlets	2703	68	40	61	0.20
Wandsworth	2124	60	35	27	0.29
Westminster	1336	39	34	39	0.33

in segregation across areas. We provide two cases. First, we compare two similar pairs of LAs, showing that quite small differences in their characteristics imply different outcomes of inference; these are North-East Lincolnshire and North Lincolnshire, and Blackburn and Oldham. Second, we compare all the different LAs in inner-city London, and consider which pairwise comparisons yield significant differences. Table 8 shows the descriptive statistics and the dissimilarity indices of the LAs. North-East Lincolnshire and North Lincolnshire have almost the same number of pupils, 2005 and 2011 respectively, but differ in the number of schools, 46 vs. 57 respectively, and consequently also average cohort size. They differ as well in the percentages of children eligible for free school meals, 21% vs. 13%. The dissimilarity index for North-East Lincolnshire is 0.43, higher than that of North Lincolnshire, which has an index of 0.36. Blackburn and Oldham differ rather more in size, but have closer average unit sizes, and slightly higher percentages of children eligible for free school meals.

Are the school allocations in North-East Lincolnshire more segregated than those in North Lincolnshire? Table 9 shows that the observed D marginally overstates the level of segregation in each local authority, but the bias-corrected estimates of D_{pop} do not alter the ranking. Table 9 further presents the various test procedures and confidence intervals as described in the previous section. Here we generate 999 bootstrap samples. The LR test for no systematic segregation clearly rejects for both LAs, with both bootstrap P values equal to 0. The rejection of the null of equal segregation in North-East Lincolnshire and North Lincolnshire depends on the test statistics employed. Using the test statistics W_{dc} and T_{pb} we reject the null of equal segregation in the two LAs at the 5% and 1% level respectively.

Table 9: Bias-corrected dissimilarity indices, confidence intervals and test statistics for North-East and North Lincolnshire

	North-East Lincolnshire	North Lincolnshire
D	0.433	0.364
D_{bc}	0.420	0.322
D_{dc}	0.416	0.334
LR-test, boot. P value	0	0
CI-W	[0.386-0.481]	[0.306-0.421]
CI-W_{pb}	[0.380-0.487]	[0.275-0.452]
CI-T_{pb}	[0.371-0.466]	[0.265-0.371]
CI-W_{dc}	[0.367-0.465]	[0.278-0.390]
	$H_0 : D_{\text{pop,NEL}} = D_{\text{pop,NL}}$, P values	
W	0.067	
W_{pb}	0.114	
T_{pb}	0.000	
W_{dc}	0.032	

Notes: CI are 95% confidence intervals. Number of bootstrap repetitions 999.

Table 10 shows test statistics for Blackburn and Oldham. In this example, we can reject, with a high degree of confidence, the null of equal segregation in these areas. This greater confidence than in the Lincolnshire example is possible, despite similar segregation levels, because the local authorities are slightly larger and the minority proportions are higher.
For our second illustration, Table 11 compares observed and density-corrected segregation levels across the 12 local authorities in Inner London. The density correction makes little differences to the ranking of segregation levels, with just Wandsworth and Greenwich switching positions. The test statistics show that the LAs can be approximately divided into three groups, with possible multiple membership, where the tests do not reject the null of equal segregation. These groups are: Tower Hamlets, Southwark and Hackney, with the lowest level of segregation; Hackney, Camden and Lambeth, with medium level of segregation; and Wandsworth, Greenwich, Hammersmith & Fulham, Lewisham, Kensington & Chelsea and Westminster with the highest level of segregation. Islington is a medium segregation LA with some overlap with the group of highest segregation LAs.

Table 10: Bias-corrected dissimilarity indices, confidence intervals and test statistics for Blackburn and Oldham

	Blackburn	Oldham
D	0.342	0.472
D_{dc}	0.306	0.446
LR-test, boot. P value	0	0
CI-T_{pb}	[0.288-0.362]	[0.420-0.485]
CI-W_{dc}	[0.263-0.348]	[0.410-0.483]
	$H_0 : D_{\text{pop,Blackburn}} = D_{\text{pop,Oldham}}$, P values	
T_{pb}	0.000	
W_{dc}	0.000	

Notes: CI are 95% confidence intervals. Number of bootstrap repetitions 999.

Table 11: Bias-corrected dissimilarity indices, confidence intervals and test statistics for Inner London

	D	D_{dc}	LR (P)	CI-T_{pb} / CI-W_{dc}	P values for tests of equivalence of D_{pop}: Sout	Hack	Camd	Lamb	Isli	Wand	Gree	Hamm	Lewi	Kens	West
Tower Hamlets	.197	.162	0	[.126-.192]	.725	.202	.176	.040	.014	.003	.000	.001	.000	.000	.000
				[.125-.198]	.899	.426	.389	.081	.016	.005	.001	.003	.000	.000	.000
Southwark	.206	.165	0	[.137-.201]		.368	.310	.094	.030	.000	.000	.000	.000	.000	.000
				[.128-.202]		.502	.454	.107	.022	.007	.001	.004	.000	.001	.000
Hackney	.219	.184	0	[.154-.225]			.875	.466	.234	.016	.000	.006	.000	.002	.000
				[.142-.226]			.897	.388	.123	.054	.019	.024	.002	.005	.000
Camden	.231	.188	0	[.153-.236]				.631	.338	.020	.016	.008	.002	.004	.000
				[.140-.237]				.511	.195	.098	.045	.044	.007	.009	.001
Lambeth	.240	.209	0	[.172-.241]					.555	.046	.034	.028	.002	.014	.002
				[.170-.248]					.457	.250	.128	.111	.020	.025	.004
Islington	.257	.231	0	[.183-.258]						.214	.126	.098	.038	.064	.008
				[.188-.273]						.690	.489	.351	.142	.101	.033
Wandsworth	.290	.243	0	[.219-.292]							.853	.587	.376	.314	.124
				[.200-.286]							.795	.561	.300	.187	.081
Greenwich	.286	.251	0	[.226-.291]								.655	.494	.350	.150
				[.213-.288]								.696	.398	.239	.106
Hammersmith & Fulham	.303	.264	0	[.226-.323]									.911	.653	.396
				[.208-.319]									.776	.467	.318
Lewisham	.304	.274	0	[.244-.312]										.663	.388
				[.235-.312]										.571	.382
Kensington & Chelsea	.317	.296	0	[.231-.347]											.779
				[.230-.361]											.880
Westminster	.328	.302	0	[.257-.347]											
				[.252-.352]											

Notes: CI are 95% confidence intervals. Number of bootstrap repetitions 999.

9 Conclusions

To make statements about the true underlying degree of segregation, or understand the processes causing segregation, it is desirable to measure the level of systematic segregation. However, where minority proportions and unit sizes are small, the level of segregation observed by researchers in their data is known to be significantly greater than systematic segregation. Furthermore, because the size of the bias of observed segregation over systematic segregation is known to be a function of minority proportion, unit sizes and systematic segregation, differences in any of these parameters between areas or over time may lead to incorrect inferences.

In this paper we have proposed and tested procedures for adjusting the dissimilarity index of segregation for this bias. Our corrections work well provided both the minority proportion and unit size are not very small. Where very small minority proportions and unit sizes render our corrections useless, we show that levels of segregation are often not statistically distinguishable from zero. We have developed and tested our statistical framework using the index of dissimilarity, D, but it can, in principle, be extended to other segregation indices.

From our statistical framework we have developed tests for a null of no systematic segregation; a null of equal segregation in two areas; and have established confidence intervals for levels of systematic segregation. In tests using unit sizes, minority proportions and underlying segregation levels similar to those encountered by social scientists, the Wald statistics using the bootstrap variance estimate for the bias-corrected estimators and the test based on the equal-tail bootstrap P value for the t test (T_{pb}) are found to perform best. The methods proposed in this paper provide a framework for more reliable inference as to levels of segregation, which will aid the further investigation of the causes of segregation.

Appendix

In this Appendix, we consider two methods of bias correction, which, although somewhat effective, turn out not to be as much so as the density-correction method of Section 3.

The expectation of the folded normal variable Z of which the density is given by (5) is easily seen to be $\mathrm{E}(Z) = 2\phi(\theta) + \theta\bigl[2\Phi(\theta) - 1\bigr] \equiv m(\theta)$. If we think of Z as an estimator of θ, then the bias is

$$b(\theta) = m(\theta) - \theta = 2\bigl[\phi(\theta) + \theta(\Phi(\theta) - 1)\bigr].$$

The bias function $b(\theta)$ is graphed in Figure 6. It decreases monotonically from its value of $(2/\pi)^{1/2}$ at $\theta = 0$, which corresponds to $\mu = 0$, and tends rapidly to 0 for values of θ greater than around 2.5.

Recall that $\hat{\theta}_j = |\hat{p}_j^1 - \hat{p}_j^0|/\hat{\sigma}_j$. Using the function $b(\theta)$ to estimate the bias of $\hat{\theta}_j$ leads to a bias-corrected estimator of D_{pop}:

$$D_{bc,1} = D - \frac{1}{2}\sum_{j=1}^{J} \hat{\sigma}_j b(\hat{\theta}_j),$$

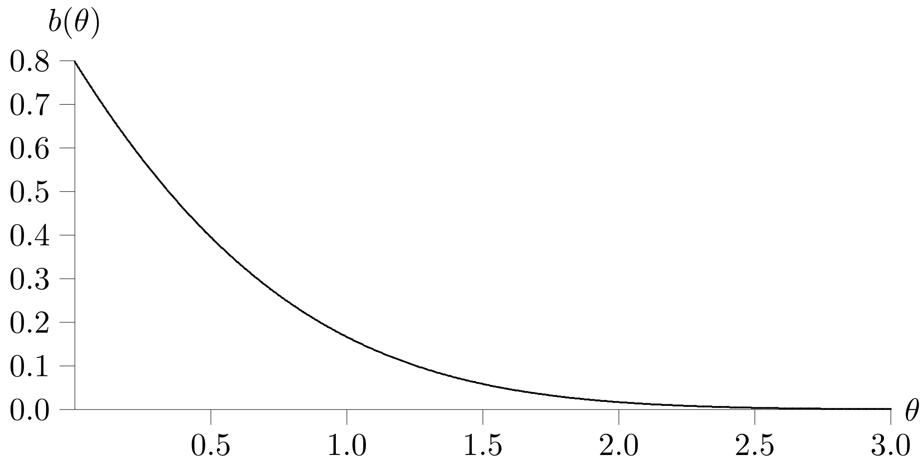

Figure 6: Bias function $b(\theta_j)$

As with the bootstrap bias correction, because of the shape of $b(\theta)$, we don't expect this correction to work well with small unit sizes combined with small values for D_{pop}.
Another approach pretends that $\tilde{\theta}_j$ really has expectation $m(\theta_j)$:

$$\mathrm{E}\big[\tilde{\theta}_j - m(\theta_j)\big] = 0.$$

We can treat this relation as an estimating equation for θ_j, thereby defining a new estimator $\hat{\theta}_j^{bc}$ as

$$\hat{\theta}_j^{bc} = m^{-1}(\tilde{\theta}_j).$$

Since in practice we must estimate σ_j, we end up with the bias-corrected estimator

$$D_{bc,2} = \frac{1}{2}\sum_{j=1}^{J} \hat{\sigma}_j m^{-1}\Big(\max\big[(2/\pi)^{1/2}, \hat{\theta}_j\big]\Big).$$

The inverse function m^{-1} cannot be expressed analytically in closed form, but it is easy to compute. Its argument must not be smaller than $(2/\pi)^{1/2}$, since that is the smallest value of $m(\theta_j)$. Thus any $\hat{\theta}_j$ smaller than this cut-off leads to a zero contribution to $D_{bc,2}$.
It is clear that the new estimator $D_{bc,2}$ is still biased, for two reasons. First, m is a nonlinear function and, second, the random quantities $\hat{\sigma}_j$ appear in the denominator of the argument of m^{-1}.

References

Allen, R. and Vignoles, A. (2007), What Should an Index of School Segregation Measure? *Oxford Review of Education*, 33, 643-668.

Boisso, D., Hayes, K., Hirschberg, J. and Silber, J. (1994), Occupational Segregation in the Multidimensional Case, *Journal of Econometrics*, 61, 161-171.

Burgess, S., McConnell, B., Propper, C. and Wilson, D. (2006), The Impact of School Choice on Sorting by Ability and Socio-economic Factors in English Secondary Education, in: L. Woessmann and P. Peterson (eds), *Schools and the Equal Opportunity Problem*, MIT Press, Cambridge.

Carrington, W. J. and Troske, K. R. (1997), On Measuring Segregation in Samples with Small Units, *Journal of Business and Economic Statistics*, 15, 402-409.

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

Davidson, R. (2009), Reliable Inference for the Gini Index, *Journal of Econometrics*, 150, 30-40.

Davidson, R. and MacKinnon, J.G. (2000), Bootstrap Tests: How many Bootstraps? *Econometric Reviews*, 19, 55-68.

Davison, A.C. and Hinkley, D.V. (1997), *Bootstrap Methods and their Applications*, Cambridge University Press, Cambridge.

Duncan, O. and Duncan, B. (1955), A Methodological Analysis of Segregation Indexes, *American Sociological Review*, 20, 210-217.

Falk, R. F., Cortese, C. F. and Cohen, J. (1978), Utilizing Standardized Indices of Residential Segregation: Comment on Winship, *Social Forces*, 57, 713-716.

Gibbons, S. and Telhaj, S. (2006), Are Schools Drifting Apart? Intake Stratification in English Secondary Schools, CEEDP 64, Centre for the Economics of Education, LSE.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.

d'Haultfœuille, X. and Rathelot, R. (2011), Measuring Segregation on Small Units: A Partial Identification Analysis, CREST Working Paper 2011/18.

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

Leone, F.C., Nelson, L.S. and Nottingham, R.B. (1961), The Folded Normal Distribution, *Technometrics*, 3, 543-550.

Massey, D.S. and Denton, N.A. (1988), The Dimensions of Residential Segregation *Social Forces*, 67, 281-315.

Mora, R. and Ruiz-Castillo, J. (2007), The Invariance Properties of the Mutual Information Index of Multigroup Segregation, Working Paper 07-75, Department of Economics, Universidad Carlos III de Madrid.

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

Rathelot, R. (2012), Measuring Segregation when Units are Small: A Parametric Approach, *Journal of Business and Economic Statistics*, 30, 546-553.

Söderström, M. and Uusitalo, R. (2010), School Choice and Segregation: Evidence from an Admission Reform, *The Scandinavian Journal of Economics*, 112, 55-76.

White, M. J. (1986), Segregation and Diversity Measures in Population Distribution, *Population Index*, 52, 198-221.

Winship, C. (1977), A Revaluation of Indexes of Residential Segregation, *Social Forces*, 55, 1059-1066.

Winship, C. (1978), The Desirability of Using the Index of Dissimilarity or any Adjustment of it for Measuring Segregation: Reply to Falk, Cortse and Cohen, *Social Forces*, 57, 717-720.