

Measuring conditional segregation: methods and empirical examples^{*}

by

Olof Åslund^{**} and Oskar Nordström Skans^{***}

May 19, 2005

Abstract

In empirical studies of segregation it is often desirable to quantify segregation that cannot be explained by underlying characteristics. To this end, we propose a fully non-parametric method for accounting for covariates in any measure of segregation. The basic idea is that given a set of discrete characteristics, there is a certain probability that a person belongs to a particular group, which can be used to compute an expected level of segregation. We also demonstrate that a modified index of exposure has both favorable analytical features and interpretational advantages in such settings. The methods are illustrated by an application to ethnic workplace segregation in Sweden. We also show how one can use a measure of exposure to study the earnings consequences of segregation stemming from different sources.

Keywords: exposure, covariates, ethnic workplace segregation

JEL codes: C15, J15, J42

^{*} We are grateful for comments from Per Johansson, Eva Mörk, Roope Uusitalo, seminar participants at IFAU and Växjö University. Åslund acknowledges financial support from Jan Walander's and Tom Hedelius' foundation. Order of authors is according to the English alphabet and not related to contribution

^{**} IFAU, P.O. Box 513, SE-751 20 Uppsala, Sweden, +46 18 471 70 89, olof.aslund@ifau.uu.se

^{***} IFAU, +46 18 471 70 79, oskar.nordstrom-skans@ifau.uu.se.

Table of contents

1	Introduction	3
2	Segregation conditional on observed characteristics	5
3	Generalizations	9
3.1	Exposure in the multigroup case	9
3.2	Other measures of segregation	11
3.3	Metrics of “excess” exposure	12
4	An empirical application	14
4.1	Data.....	14
4.2	Expected and excess exposure.....	16
4.3	Other measures of segregation—randomization results	19
4.4	Exposure and earnings.....	20
5	Concluding remarks.....	24
	References.....	26
	Appendix.....	27
	The expected value of the Duncan index.....	27
	Definitions of the Duncan index and the Gini coefficient	28

1 Introduction

The issue of segregation, i.e. inequality in the distribution of people across units (neighborhoods, workplaces, schools), has generated a vast literature in many fields of social science. Numerous empirical studies on residential and workplace segregation between races or genders have been accompanied by methodological work investigating properties of various measures of segregation. There are basically three types of questions that could be addressed by empirical studies of segregation: patterns, sources and consequences.

A majority of the previous empirical investigations concern patterns; the typical study investigates black-white residential segregation in the US. Detecting the sources of segregation has been the topic of a smaller methodological and empirical literature. In such a context one often wants to separate e.g. the ethnic workplace segregation that comes from the fact that education may vary across ethnic groups, from workplace segregation that is in a sense “purely” ethnic. The consequences of segregation, in turn, are a natural link between segregation studies and other strands of economics. For example, in recent years economists have shown a growing interest in the role of social contacts and social interactions, which is clearly linked to the issue of segregation.¹ In this context, segregation can be seen more as a factor affecting e.g. earnings and employment, rather than an outcome in its own right.

The first contribution of this paper is a method for adjusting any measure of segregation for the characteristics of the population. Our method is more general and flexible than the methods used in previous studies. Second, we show that an index of exposure that excludes the individual him-/herself in the calculation has both interpretational advantages and nice analytical features in terms of accounting for covariates when measuring segregation. The exposure index has the characteristics of each individual’s contacts as its starting point, which highlights the connection to the literature on social interactions.

We also illustrate the proposed methods by an analysis of ethnic workplace segregation in Sweden. First, we compare unconditional and conditional measures of exposure and segregation. Then, we exemplify the link between segregation and peer effects at the individual level, by studying the correlation between wages and own-group exposure. The empirical investigation uses

¹ Echenique & Fryer (2005) argue along similar lines.

matched employer-employee data covering the entire Swedish working-age population in 2000.

There are basically two groups of non-spatial segregation indices (see Massey & Denton 1988): measures of evenness and measures of exposure. The first group measures deviations from an even distribution of groups across units (but can be corrected to measure the deviation from a random distribution—see the discussion below). Traditional indices include the Gini coefficient, the Duncan index and entropy indices such as the Theil index. While these measures have been widely used, there are reasons to give more attention to indices of exposure. These indices, such as the index of isolation, quantify the exposure that members of a given group have to some group (own or other). In studies of non-market interactions between individuals, indices of exposure are suitable since they provide information on the number of contacts a person has with people carrying certain characteristics.

Several papers have described the characteristics of existing measures of segregation and/or developed new measures (e.g. Massey & Denton 1988, Flückiger & Silber 1999, and Hutchens 2004). Much of the research has focused on segregation between two groups. However, stressing the increased ethnic diversity in many societies, some authors have argued for the use of multigroup segregation measures (Boisso et al. 1994, Reardon & Firebaugh 2002).

Carrington & Troske (1997) make an important methodological point. They argue for the use of randomness rather than evenness as benchmark in studies of segregation. The key insight is that also random allocation will result in non-negligible segregation if units are small. They therefore relate segregation in actual distributions to the segregation observed in simulated distributions. We extend their work by showing how to control for observed individual and unit-related characteristics in creating expected (or “counterfactual”) distributions.

Our method for creating a conditional counterfactual distribution is non-parametric and thus accounts for all interactions between the factors considered. The factors conditioned on are allowed to vary both between and within units. The basic idea is that for each combination of characteristics there is a probability that an individual belongs to a particular group. For a given distribution of individuals over units (regardless of group), we can then assign a probability that an individual belongs to this group. The measure of “expected segregation” can then be calculated analytically or simulated, depending on the

measure of segregation and which type of variables one wishes to condition on. The method is easily extended to cases with several different groups.

Some previous studies use related empirical approaches to identify the sources of segregation. Bayer et al. (2004) investigate how differences across groups in sociodemographic factors explain residential segregation using a method based on linear regressions. Reardon et al. (2000) use modified versions of the Theil index to decompose school segregation into geographic and racial components. Kalter (2000) links the Duncan index to the multinomial logit model to take covariates into account in measures of inequality. Söderström & Uusitalo (2005) analyze segregation effects of a school choice reform in Sweden, using a linear regression model accounting for observed characteristics. Although different, the methods used in these studies are all limited in the sense that they are not valid for all segregation indices, and/or make restrictive assumptions regarding the data generating process.

The rest of the paper is outlined as follows. Section 2 presents methods for computing segregation conditional on observed characteristics, using a modified index of exposure in a two-group situation. We then generalize the method to the multi-group case, and show that its logic can be applied to any measure of segregation. A discussion of how to define “excess” exposure concludes section 3. We begin section 4 by describing the Swedish linked employer-employee data. The empirical applications in the following parts of the section first show how conditioning on various sets of characteristics affects different measures of segregation. Then we demonstrate the link between segregation and social interactions by studying the correlation between own-group exposure and earnings. Section 5 concludes.

2 Segregation conditional on observed characteristics

In general, the concept of segregation aims to capture systematic sorting over units (e.g. geographical areas, schools or workplaces) by individuals belonging to different groups (defined by for example gender or ethnicity). Thus, segregation can be said to occur if the distribution of groups over units is significantly

different from what would result from a random allocation.² Below we first define a measure of exposure that has randomness as its natural benchmark. This measure also has the advantage of being linear, which enables us to calculate exposure without simulations, and at the individual level. We then go on to show how to test for random allocation conditional on observed characteristics.

One way to think of segregation that explicitly incorporates randomness as the baseline is to ask whether (e.g.) my own minority status predicts the minority status of the people around me (e.g. at my workplace). Following this line of thought, we define segregation to occur if an individual is more “exposed” to his own group than he would be if the distribution of individuals across units was random. By exposure we mean the average group characteristics of *other* people within the same unit. We therefore exclude the individual himself in the calculations.

Now, consider the case with two groups, where some individuals belong to a minority (m , $D^m = 1$) and others do not ($D^m = 0$). There are N individuals in total, whereof N^m belongs to the minority. Suppose that a minority and a majority individual “picked” their unit peers in a random process. The expected peer fraction for both individuals is then equal to the minority fraction in the population, N^m / N .³ Thus, with this individual-centered approach, measures of intra- and intergroup exposure are directly comparable.⁴

To formalize this argument, assume that there is a set of units denoted by w , each of size s^w . We define individual i 's *exposure* (e) to a minority within his unit $w(i)$ as the fraction of *others* in the unit that belong to the minority:

$$e_{i,w(i)} = \frac{1}{s^{w(i)} - 1} \sum_{\substack{w(j)=w(i) \\ j \neq i}} D_j^m . \quad (1)$$

² Yet, many standard measures of segregation assess how far an observed distribution is from evenness rather than randomness (see e.g. Carrington & Troske 1997, for a discussion).

³ Technically, this is an approximation since we do not exclude the individual i from the minority fraction. If we did, it should be $\frac{N^m-1}{N-1}$ for minority individuals and $\frac{N^m}{N-1}$ for non-minority individuals. Obviously, this correction plays no role whatsoever at any realistic (minority) population size

⁴ However, from a bird's perspective, minority individuals will on average be in units with higher minority representation (since a minority individual is always exposed to himself). With such an approach, the logic of section 3.2 should be applied to measure conditional segregation.

Further, let ε^m be the *average exposure* to minority individuals by individuals who belong to the minority, which is given by:

$$\varepsilon^m = \frac{1}{N^m} \sum_{i:D_i^m=1} e_i . \quad (2)$$

Since majority individuals should be as exposed to minorities as minority individuals (in the absence of segregation), using E to denote the expected values in the absence of segregation gives:

$$E(e_{i,w(i)} | D_i^m = 1) = E(e_{i,w(i)} | D_i^m = 0) \quad (3)$$

Equation (3) states that an individual's exposure to minorities does not depend on his minority status (if there is no segregation). It naturally follows that the expected exposure for both groups should equal the fraction of minority individuals in the population. Thus, for the minority group we get:

$$E(\varepsilon^m) \equiv E(e_{i,w(i)} | D_i^m = 1) = \frac{N^m}{N} \quad (4)$$

Testing whether the actual average calculated according to (2) is equal to the expected “non-segregation” value given by (4) is a test for segregation.⁵

So far we have only described how to test for segregation in the unconditional sense. In a second step we would like to test the hypothesis that minorities are only exposed to minorities to an extent that is motivated by the distribution of some observed characteristics X . In other words, we want to test whether minority individuals' exposure to minorities can be explained by the characteristics of their units, or the people who belong to the unit. For example we might want to study whether immigrants work with immigrants because many immigrants tend to live in urban areas, or because people with the same educational background tend to work together.

⁵ Note however that when computing standard errors for statistical tests, we need to account for the fact that the observations are not independent within units (within a unit e_i is always equal for all individuals with the same minority status). Thus, we use unit level cluster-corrected standard errors in our empirical application below.

In this case we wish to calculate expected exposure in the absence of segregation *above* what can be explained by the X -characteristics. We put no restrictions on the distribution of characteristics; X may vary both within units and between units. The method is completely non-parametric for discrete X 's.

For each unique realization (x) of the vector X we calculate the fraction of minority individuals and denote it by p . For an arbitrary individual j 's realization we get:

$$p(x_j) = \frac{N^m(X = x_j)}{N(X = x_j)} \quad (5)$$

where $p(x_j)$ can be interpreted as the probability that an individual with characteristics x_j belongs to the minority. For each individual we can calculate the expected exposure conditional on the distribution of X using (5) for the other people in the unit:

$$E(e_i | x^j_{\forall j:w(j)=w(i), j \neq i}) = \frac{1}{s^{w(i)} - 1} \sum_{\substack{w(j)=w(i) \\ j \neq i}} p(x_j) \quad (6)$$

Furthermore, taking the average over all individuals belonging to the minority we get

$$E(\varepsilon^m | x^j_{\forall j:w(j)=w(i), j \neq i}) = \frac{1}{N^m} \sum_{i:D_i^m=1} \frac{1}{s^{w(i)} - 1} \sum_{\substack{w(j)=w(i) \\ j \neq i}} p(x_j) \quad (7)$$

Note that when all individuals have the same probability (N^m/N), (7) collapses to (4).

By comparing (2) to (7) we will get a test for segregation above what can be explained by the distribution of the underlying characteristics X . Empirical examples are presented in section 4 below. Note that the derivation of (7) crucially hinges on the fact that ε^m is a linear function of the “peers” values of D^m , making it straightforward to replace D^m with the probability $p(x)$. For non-linear measures, such as when the individual i himself is included in the

measure of exposure, one is limited to using the general techniques presented in section 3.2 below.

Note also that the test is based on using expression (5) to predict the expected minority status of each individual in the absence of segregation. One could also use e.g. “propensity scores” from binary regression models to predict the minority status of individuals. However, since such parametric models would impose unnecessary restrictions on the data generating process we will apply the non-parametric method described by (5) throughout this paper.⁶

3 Generalizations

We begin this section by generalizing the method presented above to the multi-group case. We then show how the logic can be applied to widely used measures of segregation, and the limitations to such applications. The section also discusses different metrics of the amounts of segregation and their economic interpretations.

3.1 Exposure in the multigroup case

So far we have assumed that there is only one minority group. In many applications we may have a number of parallel groups, or construct such by combining different dichotomous variables (e.g. gender and race). Such cases can easily be studied using the framework presented above without changing the interpretation. Below we will discuss a straightforward generalization. For convenience we will focus on own-group exposure, rather than cross-exposure.⁷ It should be noted however, that adapting the tools described below to an analysis of cross-exposure is straightforward.

Defining a set of groups $g = (1, \dots, G)$ we can measure individual i 's exposure to group g as before (in eq. 1):

⁶ Parametric methods may be preferred when there are only a few observations per cell (possibly as a result of essentially continuous variables).

⁷ Interesting measures of cross-exposure include e.g. the exposure of Iraqi immigrants to other Arab immigrants (to test for segregation according to language or religion), and the exposure of white men to black women (to test for interaction effects in segregation between gender and race).

$$e_{i,w(i)}^g = \frac{1}{s^{w(i)} - 1} \sum_{\substack{w(j)=w(i) \\ j \neq i}} D_j^g. \quad (8)$$

As before we can also take the average over all individuals, or a subset of individuals belonging to a set of groups $g \in \Gamma$, to calculate average *own-group exposure*:

$$\mathcal{E}_\Gamma^g = \frac{1}{\sum_{g \in \Gamma} N^g} \sum_{i: g(i) \in \Gamma} e_i^g D_i^g \quad (9)$$

It is worth noting that we study the average fraction of other individuals within the unit that belongs to the same group in order to reduce the dimensionality of the problem. Because of this, we may use the same strategy as in (5) above when calculating the expected distribution without segregation: We first calculate the fraction of individuals belonging to group g for each value of X (without excluding the probability that one X contains all individuals) and denote it by p . Then, for an arbitrary individual j :s realization x we have a p corresponding to g given by:

$$p^g(x_j) = \frac{N^g(X = x_j)}{N(X = x_j)} \quad (10)$$

Thus, as before we get the conditional expected exposure:

$$E(\mathcal{E}_\Gamma^g | x^j_{\forall j: w(j)=w(i), j \neq i}) = \frac{1}{\sum_{g \in \Gamma} N^g} \sum_{i: g(i) \in \Gamma} \frac{1}{s^{w(i)} - 1} \sum_{\substack{w(j)=w(i) \\ j \neq i}} p^{g(i)}(x_j) \quad (11)$$

The actual own-group exposure described by (9) measures the fraction of others in one's own unit that belong to the same group as oneself, calculated over all individuals belonging to a group in Γ . By contrast, the expected own-group exposure described by (11) measures the fraction of individuals in my own unit that *are expected to* belong to the same group as myself. In our empirical setting these measures are the average—observed or expected—

fractions of coworkers with the same country of origin as the individual, calculated as an average over all immigrants.

3.2 Other measures of segregation

The method as described so far crucially hinges on the linearity of the measure of exposure. In this subsection we describe how the logic can be applied to other measures of segregation.

Carrington & Troske (1997) show how randomly allocating minority status, keeping the size distribution of units as given, gives a reasonable baseline for measuring segregation irrespective of the applied measure. Their strategy basically amounts to first calculate a measure of segregation Z , then randomly allocate minority status and calculate a counterfactual measure $E(Z)$. By contrasting Z and $E(Z)$, measures of systematic sorting are obtained.

The logic underlying our methodology can be used to extend the method of Carrington & Troske (1997) in order to facilitate an analysis of conditional segregation using any measure of segregation. In order to measure segregation conditional on X , we need to obtain a measure of segregation that is based on an allocation of minority status that is random conditional on the covariates: $E(Z | X)$. Conceptually, we are interested in calculating the values of segregation given that workers were randomly allocated holding the sizes of the units *and* the distribution of X -variables over units constant. Our methodology as explained in section 2 has the advantage of being linear, which gives the opportunity to calculate expected exposure directly and at the individual level. When using the logic behind it for non-linear measures of segregation we generally have to follow Carrington & Troske and rely on simulations.⁸

Our strategy is to achieve a counterfactual distribution by randomly allocating minority status to individuals within each cell defined by a specific realization x , using a *probability* equal to the fraction of minority individuals given by equation (5). We can then proceed by calculating any index of segregation using the conditional random allocation of minority status.⁹ For each individual we allocate a continuous random variable v , which is uniformly distributed be-

⁸ For some measures, such as the Duncan index, that are linear in the units it is also possible to calculate the expected values directly in cases where the conditioning variables are fixed for each unit. This procedure is presented in the appendix.

⁹ The empirical examples of section 4.3 suggest that the randomization procedure yields results that are practically identical to the expected values that are calculated directly when studying exposure.

tween 0 and 1. Then we infer a counterfactual minority status q based on the relationship between v and $p(x)$ using (5):

$$\begin{aligned} q_i &= 0 \text{ if } v > p(x_i) \\ q_i &= 1 \text{ if } v \leq p(x_i) \end{aligned} \tag{12}$$

We can then calculate counterfactual measures of segregation $E(Z|X)$ based on the distribution of q , and contrast it to the actual measure Z . Since replicated simulations provide measures of uncertainty, we also get tools for inference.

This methodology has the feature that it is applicable to any measure of segregation. Moreover, the methodology can, just as in section 2, be used to condition on any number of characteristics and all types of (discrete) characteristics, either unit based and/or individual based (i.e. some or all characteristics may vary within units). Although the exposition above is limited to a two-group case (minority vs. not minority), it is straightforward to generalize the logic to multi-group cases by making the assignment in (12) group specific.

While this logic is both general and conceptually straightforward, some of the advantages of using the index of exposure are lost. First, the general strategy relies on repeated simulations, which is cumbersome when looking at very large datasets or when analyzing many sub-samples. Second, the measure of exposure defined in section 2 has the property that both its actual and expected values are defined at the individual level. Thus, measuring segregation by exposure as defined in section 2 allows us to study *who* deviates from the expected pattern, which is convenient when studying the consequences of segregation (as in section 4.3 below).

3.3 Metrics of “excess” exposure

The presented methods provide an opportunity to test whether the observed level of exposure is statistically different from what one would find if the distribution of people was random conditional on the covariates. For an economic interpretation one would often like to relate actual exposure to expected exposure in order to get a measure of “excess” exposure. Clearly, however, there is no definition that suits all types of applications. Focusing on exposure as an outcome typically requires other modifications than when we think of exposure as an explanatory factor to another variable. Thus, our view is that there is no need to confine oneself to one “all purpose” definition.

There are obviously many ways of defining excess exposure (or excess segregation). Carrington & Troske (1997) suggest using an “index of systematic segregation” of the form

$$\hat{Z} = \frac{Z - E(Z)}{1 - E(Z)} \quad (13)$$

(for $Z \geq E(Z)$), i.e. the difference between actual segregation and expected segregation over the maximum value of non-random segregation. When this formula is used with the measure of exposure discussed in section 2, it corresponds to the fraction of minority individuals that have to be completely segregated, given that all other individuals are subject to the expected level of exposure.¹⁰

An appealing alternative is to simply divide actual exposure by the expected exposure rate:

$$R_X^m = \frac{\varepsilon^m}{E(\varepsilon^m | X)} \quad (14)$$

This *relative overexposure* is very easily interpreted: “The average immigrant has R_X^m times as many immigrant coworkers compared to what we would expect if the distribution was random, given X ”. Another advantage is that this fraction is directly comparable in analyses on subgroups. One may, e.g., wish to compare excess exposure by region of residence, industry, ethnicity or gender. Suppose we study two regions, and that the immigrant fraction and thereby the expected exposure (ignoring differences in the distribution of X) is twice as high in one region compared to the other. If actual exposure also differs by a factor of two, it is intuitive to think of excess segregation as being the same in the two regions. This is what R_X tells us.

The methods outlined in sections 2 and 3 provide a way to condition on observed characteristics in the measurement of segregation. We have shown that the basic method can be applied to any measure of segregation, both in two-group and multigroup cases. Furthermore, using exposure excluding the individual himself has both analytical and interpretational advantages. In the next

¹⁰ Assume that f is the fraction of the minority that is only exposed to their own group. Then it must be that $\varepsilon^m = f * 1 + (1 - f)E(\varepsilon^m)$. Thus, $f = (\varepsilon^m - E(\varepsilon^m)) / (1 - E(\varepsilon^m))$.

section we apply these methods in an empirical analysis of immigrant-native workplace segregation in Sweden.

4 An empirical application

This section demonstrates the consequences of conditioning on observed characteristics in the measurement of segregation in the context of ethnic workplace segregation in Sweden. We present measures of expected and excess exposure using the methods for calculating expected values presented above. Then we present conditional results based on randomization for two commonly used measures of segregation: the (Duncan) index of dissimilarity and the Gini coefficient. The section is concluded by an analysis of the connection between exposure and earnings at the individual level. We begin, however, with a short description of the data.

4.1 Data

The data used in this paper is a linked employer-employee data set, the IFAU database, covering the entire Swedish economy in 2000. The data are based on tax records and contain annual information on all 16–65 year-old employees receiving remuneration from Swedish employers (both private and public). We focus primarily on segregation between immigrants and natives. Immigration status is measured by a grouped variable containing country or region of birth.¹¹

We use tax-record earnings information to construct the employment status of workers. The earnings data contain annual earnings, and the first and last remunerated months from a specific employer. From this we construct a measure of monthly earnings for all employment spells that cover the month of No-

¹¹ We wish to consider workers as immigrants if they are born abroad, *excluding* adoptees. The reason for this restriction is that we are interested in workplace sorting according to ethnic dimensions which should be a function of the foster parents rather than the biological parents for children that were adopted at a very young age. In practice, the workers we consider as adopted, and thus code as Swedish-born, are i) born outside of Sweden, ii) arrived to Sweden before age 3 and iii) have the country of birth of *both* their parents coded as missing. Some children arriving with relatives instead of biological parents could be miscoded as Swedes according to this procedure, but they are likely to be few.

vember.¹² We then drop all observations with average monthly earnings below 25 percent of a constructed monthly minimum wage.¹³ We only keep the job generating the highest monthly wage for each individual and year.

Table 1 Descriptive statistics.

	Immigrants	Swedish born
Age	41.8	41.0
16-29	0.151	0.209
30-49	0.572	0.496
50-65	0.277	0.295
Female	0.488	0.484
Education		
Primary or less	0.204	0.157
2-year secondary	0.269	0.313
3-year secondary	0.186	0.196
Some tertiary	0.124	0.151
At least 3 year tertiary	0.172	0.172
Graduate	0.021	0.009
Unknown	0.024	0.001
ln(monthly earnings)	9.633	9.724
Standard deviation	(0.525)	(0.515)
N	313,973	2,865,490

Note: The data cover all Swedish resident workers employed by workplaces with at least 5 employees. The workers are distributed over 115,226 workplaces in 289 municipalities and 38 industry groups.

Using the individual employment data described above, we calculate the number of employees by workplace (note that this includes the self-employed). Since we are interested in describing the composition of workers in different

¹² November is the month of measurement for Sweden's official annual employment statistics.

¹³ Swedish law does not determine a minimum wage. We define the minimum wage by the published mean monthly wage of janitors employed by local municipalities each year (14,100 SEK in 2000).

workplaces, we exclude workplaces with less than 5 employees.¹⁴ Apart from the number of employees, we characterize the workplaces by the municipality (289 groups) and industry¹⁵ (38 groups). In addition to the workplace information we use data on individual characteristics: age groups (<30, 30-49 and 50+), educational groups (7 categories) and gender. *Table 1* shows some descriptive statistics.

4.2 Expected and excess exposure

This section presents the results from an analysis of workplace segregation in Sweden using the measure of exposure presented above. We study the consequences of and arguments for conditioning on different sets of covariates in the measurement of exposure.

Row (1) of *Table 2* shows the observed level of own-group coworker exposure among immigrants in Sweden. We begin by discussing exposure to other immigrants regardless of origin, and then turn to exposure to people from one's own birth region. The actual value of "Immigrant exposure" says that the average immigrant in Sweden has about 21 percent immigrant colleagues. Row (1) also shows analytical standard errors that are cluster-corrected for dependencies within workplaces. These standard errors facilitate inference without simulations or bootstrapping, something which is not typically possible for measures of segregation.

The validity of the analytical procedure is confirmed by the fact that the results (expectations and confidence intervals) are similar to what we retrieve by simulations and bootstrapping.¹⁶ Note that in addition to the advantage of avoiding simulations, the analytical values are necessary when we, as in section 4.4, investigate the relation between earnings and expected exposure at the individual level.

The "Expected value" column in rows (2) through (6) shows the average expected value of exposure, when conditioning on various sets of variables.

¹⁴ 17.2 (18.6) percent of the native (immigrant) workers work in establishments with fewer than 5 employees. These numbers include the 5.6 percent that do not have a well-defined physical workplace.

¹⁵ Industry is based on the "reduced" 2-digit industries that are the smallest common denominator between the classification systems SNI-92 and SNI-69

¹⁶ Simulated values for expected exposure are presented in *Table 3*. Bootstrapped confidence intervals (cf. *Table 2*), based on bias-corrected results from 500 workplace-based replications were [0.2028-0.2150] and [0.0472-0.0504] respectively.

The second and fourth columns present the value of relative overexposure, R_X (i.e. observed exposure divided by the expected value).

The unconditional expected value in row (2) of just below 10 percent indicates that the observed exposure is about two times the exposure one would see under random allocation of individuals; R_X is 2.1. A first natural candidate for conditioning is human capital. Groups may differ in terms of e.g. age, gender and education, and this may be a cause of segregation in the labor market. Row (3) shows that conditioning on these variables yields a value of expected exposure that is only slightly higher than the unconditional rate.

So far we have implicitly assumed that under random allocation, a person is as likely to work in any establishment in Sweden. This is of course not reasonable in many cases; it is not surprising that workplaces in regions where there are no immigrants have few immigrant employees. A commonly used method is to restrict the analysis to a smaller region that constitutes a local labor market. However, we are often interested in an average level of segregation in e.g. a country. It is clear that conditioning on municipality increases the expected level of exposure (rows (4) and (5) in *Table 2*), indicating that the geographic distribution of immigrants and natives differ to some degree. It should, though, be noted that the geographic sorting may not be exogenous to the process of workplace segregation.

The endogeneity issue is even more important when we, as in row (6), condition also on industry. Segregation by industry can be a result of e.g. discriminatory practices that vary across industries. On the other hand, segregation in this dimension may also arise if the members of one group possess qualifications that are more suitable for a particular industry. The table gives a value of relative overexposure (R_X) of 1.3 conditional on human capital, municipality, and industry. This is certainly lower than the unconditional R_X of 2.1, suggesting that much of the segregation we observe can be explained by the factors conditioned on in the analysis. However, the results can also be interpreted as saying that even with controls for a lot (remember that the conditioning is fully interacted), immigrants still tend to have 33 percent more immigrant colleagues than what would be the case under conditionally random allocation. It is obvious that the actual value in all cases is significantly larger than the expected value, in both the economic and the statistical sense.

Table 2 Workplace segregation between immigrants and Swedish-born.

	Immigrant exposure		Region-group exposure	
(1) Actual value	0.2078		0.0489	
(standard error)	(0.0031)		(0.0008)	
[95 % confidence interval]	[0.2017–0.2139]		[0.0473–0.0505]	
Predictions	Expected value	Relative overexposure (R_X)	Expected value	Relative overexposure (R_X)
(2) Unconditional	0.0988	2.101	0.0095	5.126
(3) Conditional on human capital	0.1045	1.981	0.0104	4.702
(4) Conditional on municipality	0.1190	1.748	0.0160	3.055
(5) Conditional on human capital and municipality	0.1263	1.651	0.0180	2.711
(6) Conditional on human capital, municipality and industry	0.1561	1.333	0.0268	1.821

Note: Analytical standard errors are cluster-corrected to account for dependences within workplaces. Bootstrapped confidence intervals are bias-corrected results from 500 workplace-based replications. Human capital of colleagues is captured by gender, age (3 dummies) and education (7 dummies). Region-group exposure is immigrants’ average fraction of coworkers that are from the same region of origin (26 regions).

Turning now to region-group exposure, we find that the average immigrant has slightly less than five percent colleagues from his own birth region, which is about five times more than what we would expect from an unconditional random allocation (row (1)). It is interesting to see that conditioning on municipality decreases the relative overexposure to “countrymen” more than the corresponding measure for immigrant exposure. The value in row (4) is 40 percent lower than that in row (2) for “Region-group” overexposure, whereas the corresponding difference for immigrant exposure is 17 percent. This pattern reflects the tendency of specific immigrant groups to cluster geographically, and emphasizes the importance of conditioning to avoid premature conclusions in analyses of ethnic workplace segregation.

4.3 Other measures of segregation—randomization results

We now turn to study how our method for conditioning affects two widespread measures of segregation: the Duncan index and the Gini coefficient.¹⁷ For these measures (and any other segregation index), we can create counterfactual distributions via randomized allocation of individuals over units. These distributions are then used to compute expected levels of segregation. Note however, that this exercise is somewhat more computer intensive since it requires repeated simulations for the inference.

The structure of *Table 3* is the following: The first row of figures in each “cell” is the expected level of segregation; its standard deviation is in parentheses. The figures in brackets are the implied “indices of systematic segregation” (see section 3.3). The first column of the table presents the randomization results for immigrant exposure among immigrants. It is clear that the randomization procedure yields results that are very similar to the analytical expectations presented in *Table 2*. Thus, using the index of exposure makes simulations unneeded in this context. Columns 3 and 4 show how the Duncan index and the Gini coefficient are affected.

In their analysis of inter-firm racial segregation in Chicago, Carrington & Troske (1997) find “raw” levels of segregation that correspond roughly to the values in row (1) of *Table 3*: they report a Duncan index of 0.504 and a Gini coefficient of 0.664. It is also interesting to note that their analysis yields a degree of “systematic segregation” (\hat{Z}) that is not too far from our (unconditional) estimates: 0.251 and 0.344 respectively for Chicago compared to 0.281 and 0.411 for Sweden (row (2)).

Our analysis also shows that conditioning on an increasing number of X -variables affects the two standard indices in a way similar to what we find for the exposure index. The conditional Duncan index of row (6) gives a \hat{Z} value of 0.097. Taking the distribution across individual characteristics, municipalities and industries as exogenous, this value says that we only observe a systematic segregation that is 10 percent of what it could be. This is considerably lower than the corresponding unconditional figure of 28 percent.

¹⁷ For definitions, see the appendix.

Table 3 Randomization results: Exposure and segregation.

	Exposure	Duncan	Gini
(1) Actual value	0.208	0.432	0.602
<u>Predictions:</u>			
(2) Unconditional	0.0988 (0.0003) [0.121]	0.2094 (0.0006) [0.281]	0.3233 (0.0008) [0.411]
(3) Conditional on human capital	0.1045 (0.0003) [0.115]	0.2153 (0.0006) [0.276]	0.3294 (0.0008) [0.406]
(4) Conditional on municipality	0.1190 (0.0003) [0.101]	0.3082 (0.0007) [0.179]	0.4325 (0.0008) [0.298]
(5) Conditional on human capital and municipality	0.1263 (0.0003) [0.093]	0.3118 (0.0007) [0.174]	0.4384 (0.0008) [0.291]
(6) Conditional on human capital, municipality and industry	0.1561 (0.0003) [0.061]	0.3706 (0.0007) [0.097]	0.5184 (0.0008) [0.173]

Note: Human capital of colleagues is captured by gender, age (3 dummies) and education (7 dummies). Predictions are based on randomization (500 replications), standard deviations from these randomizations are in parentheses. Implied indices of systematic segregation (\hat{Z}) are in brackets (see equation 14 in section 3.3). See the appendix for definitions of the Duncan and Gini indices.

4.4 Exposure and earnings

We started the introduction by noting the link between segregation and other topics in economics, e.g. the study of social interactions. In this context segregation is a factor that potentially affects another variable of interest. In this subsection we illustrate one approach to study the effects of segregation on individual outcomes. It should be noted that the identification of social interaction effects is a non-trivial task, and that the analysis below abstracts from many important empirical problems that should be taken seriously in a more rigorous analysis (see e.g. Manski 1993).

We include only immigrants in this part of the analysis, and begin by estimating the following earnings equation

$$\ln y_i = \beta^M Z_i^M + \gamma e_i + u_i, \quad (16)$$

where Z_i^M contains individual characteristics (education, age, gender, dummies for region of origin, and dummies for five-year periods since migration). Individual level immigrant exposure (e_i) is the explanatory variable of primary interest. The results in column (1) of *Table 4* show that immigrants who have many immigrant colleagues have significantly lower earnings than other immigrants.¹⁸ A standard deviation in immigrant exposure amounts to 0.20,¹⁹ and according to the estimate such an increase in exposure is associated with about 6 percent lower earnings.

In (16), identification of γ comes from exposure variation in many dimensions, e.g. between municipalities and industries. An interesting issue is whether the negative correlation between exposure and earnings is present in every dimension. *Figure 1* below suggests that this is not the case. The figure plots standardized earnings²⁰ and exposure, averaged by municipality and industry respectively. The graphs tell us that earnings are actually higher in municipalities with high own-group exposure among immigrants, but that earnings fall with higher exposure when we compare industries.

In other words, changing exposure may be associated with different expectations on earnings, depending on the source of the change. A related question is whether the “effects” of expected and excess exposure differ. To this end, we run regressions where we decompose exposure into expected exposure and excess exposure in the following way:

$$e_i = (e_i - E(e_i | X)) + E(e_i | X) \quad (17)$$

¹⁸ One could worry that “immigrant dense” workplaces have more temporary jobs, meaning that the constructed monthly earnings may be misleading. For example, if a person works only in July and December, we will erroneously divide total earnings by six instead of two. As a robustness check we used total annual earnings as the dependent variable. The results were similar to those in *Table 4*, although slightly larger in absolute values.

¹⁹ The average “within-region of origin” and “within-region of origin-and-municipality” standard deviations in exposure are 0.186 and 0.160 respectively.

²⁰ The values are residuals from the model specified in (16), excluding the exposure variable.

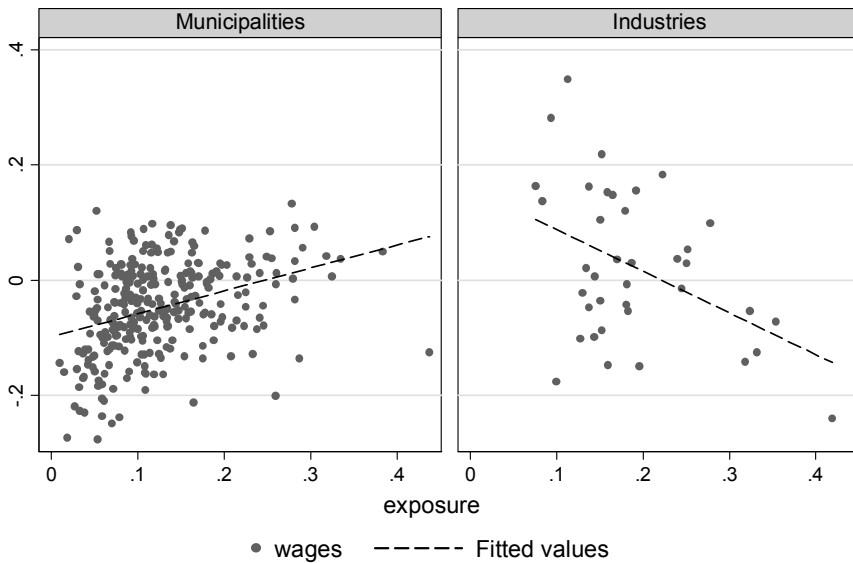
Table 4 The association between immigrant exposure and immigrant earnings.

Exposure variable	(1)	(2)	(3)	(4)
e_i	-0.291** (0.012)		-0.333** (0.013)	
$e_i - E(e_i X)$		-0.384** (0.018)		-0.335** (0.016)
$E(e_i X)$		-0.113** (0.033)		-0.327** (0.039)
Individual controls	Yes	Yes	Yes	Yes
289 municipal fe	No	No	Yes	Yes
38 industry fe	No	No	Yes	Yes
R-squared	0.24	0.24	0.31	0.31

Note: Dependent variable is logarithm of monthly earnings. All regressions include observations of 313,973 individuals in 60,068 workplaces. See equations (3) and (6) for explanations of the exposure variables. The expected exposure $E(e_i | X)$ is based on human capital, municipality, and industry. Individual control variables are education (7 dummies), age, age squared, gender, dummies for five-year periods since migration, region of origin dummies. Human capital of colleagues is captured by gender, age (3 dummies) and education (7 dummies). Standard errors (in parentheses) are cluster-corrected to account for dependencies within workplaces.

According to the estimates in column (2) of *Table 4*, earnings have a stronger negative correlation with excess exposure than with expected exposure. Both coefficients are negative and significant, but the one for excess exposure is more than three times as large in absolute value.

The next step is to see what happens to the correlation between earnings and exposure if we exploit only within-municipality / -industry variation in exposure. Indeed, this is what one would do to get closer to a causal interpretation of the estimates. Comparing columns (1) and (3) of *Table 4* suggests that the correlation between earnings and observed exposure is about the same, with and without municipality and industry fixed effects.



Notes: The regression coefficient for the municipality fitted values is 0.40 (se 0.06). The corresponding figures for industries are -0.72 (0.25).

Figure 1 Immigrant earnings and immigrant exposure by municipality and industry.

However, when these fixed effects are included in the decomposed model of column (4), there is no longer much difference between the estimates for expected and excess exposure.²¹ This is no big surprise, given the rather subtle difference in variation that the estimates are based on. Remembering the region and industry fixed effects included in the regression, we can interpret the finding in the following way. Exposure caused by the fact that one works in an industry having an unusually immigrant dense work force in my particular municipality— $(E(e_i | X))$ —correlates with earnings roughly in the same way as does excess exposure from being in a workplace that has unusually many im-

²¹ It is worth mentioning that the identification of the parameter for expected exposure comes from the interactions between human capital, municipality, and industry. Had our measure of expected exposure not been fully interacted, there would in principle be no variation given the other explanatory variables of the earnings model.

migrants *given* that it belongs to a particular industry in a particular municipality ($e_i - E(e_i | X)$).

The results above show that different variables associated with immigrant exposure have different relationships to earnings. Immigrants living in municipalities with high immigrant densities earn more than other immigrants, while immigrants in immigrant dense industries earn less. On the other hand, when we control for municipality and industry we see that the predictions based on the interactions of municipality, industry and coworkers' human capital has a negative relationship to earnings that is very similar to that of the unexplained exposure. Thus, a slightly speculative interpretation of the results is that it is as bad to work at an immigrant-dense workplace regardless of why it is immigrant-dense, *as long as* we control for the municipality (i.e. regional differences in earnings and/or exposure).

The purpose of this analysis is to illustrate how measures of exposure and segregation can be used as explanatory factors in economic analysis. It may obviously be premature to give the results a causal interpretation. Suppose, e.g., that workers are sorted across establishments based on skills, and that skills determine earnings. If immigrants are overrepresented among the low-skilled (in the unobserved sense), we would expect the correlations observed here. On the other hand it is worth noting that—in addition to individual control variables—the regressions include fixed region and industry effects, which makes a causal link more likely.

5 Concluding remarks

In the introduction we argued that there is a need to connect studies of segregation to other empirical fields in economics. To ease this process, measures of segregation need to be interpretable in an economic context. We are e.g. often interested in the factors driving segregation. In such cases there is a need for conditional measures of segregation. This paper's first contribution is a general method of conditioning on covariates in the study of segregation. The basic idea is that for any set of discrete characteristics, there is a certain probability that a person belongs to a particular group. These probabilities can then be used to compute an expected level of segregation. The method is fully non-parametric and can be used on any measure of segregation.

We also show that an index of exposure that excludes the individual him-/herself from the calculations has several advantages. It has an intuitive economic interpretation that emphasizes social contacts and is well-defined at the individual level. The measure is also generalizable to the multi-group case, and allows for straightforward cross- and subgroup comparisons. It also provides measures of expected segregation without simulations.

Our empirical application to immigrant-native workplace segregation suggests that conditioning has a substantial influence on the level of excess or systematic segregation. The qualitative impact is similar for our measure as for commonly employed measures of segregation. Still, also when controlling for—perhaps too—many factors, immigrants have disproportionately many immigrant colleagues compared what we would find with procedure based on random allocation conditional on covariates. The excess workplace segregation that remains after conditioning on e.g. region of residence comes much closer to what Bayer et al. (2004) call “pure” segregation. In many applications, this is what we would like to measure.

We also show how one can link segregation to individual labor market outcomes. The tentative estimates suggest a clear negative correlation between earnings and own-group exposure among immigrants. At face value, the results say that workers with ten percentage points higher exposure have about 3 percent lower earnings. To us, the provisional results presented here signal that further investigating the patterns and causes of ethnic workplace segregation as well as its impact on earnings and job stability are important questions for future research.

References

- Bayer P, R McMillan & KS Rueben (2004), “What Drives Residential Segregation? New Evidence Using Census Microdata”, *Journal of Urban Economics* 56, 514–535.
- Boisso D, K Hayes, J Hirschberg & J Silber (1994), “Occupational Segregation in the Multidimensional Case. Decomposition and Tests of Significance”, *Journal of Econometrics* 61, 161–171.
- Carrington WJ & KR Troske (1997), “On Measuring Segregation in Samples with Small Units”, *Journal of Business & Economics Statistics* 15 (4), 402–409.
- Echenique F & RG Fryer (2005), “On the Measurement of Segregation”, NBER working paper 11258.
- Flückiger Y & J Silber (1999), *The Measurement of Segregation in the Labor Force*, Physica-Verlag, Heidelberg.
- Hutchens R (2004), “One Measure of Segregation”, *International Economic Review* 45 (2), 555–578.
- Kalter F (2000), “Measuring Segregation and Controlling for Independent Variables” Arbeitspapiere Nr 19, 2000, Mannheimer Zentrum für Europäische Sozialforschung.
- Massey DS & NA Denton (1988), “The Dimensions of Residential Segregation”, *Social Forces* 67 (2), 281–315.
- Manski C (1993) “Identification of Endogenous Social Effects: The Reflection Problem”, *Review of Economic Studies* LX, 531–542.
- Reardon SF, JT Yun & T McNulty Eitle (2000), “The changing structure of School Segregation: Measurement and Evidence of Multiracial Metropolitan-Area School Segregation, 1989–1995”, *Demography* 37(3), 351–364.
- Reardon SF & G Firebaugh (2002), “Measures of Multigroup Segregation”, in Stolzenberg RM (ed.), *Sociological Methodology* 32, 2002, 33–67, Blackwell Publishing, Boston, MA.
- Söderström M & R Uusitalo (2005) “School choice and segregation: evidence from an admission reform”, IFAU working paper 2005:7.

Appendix

The expected value of the Duncan index

The Duncan index (D) of segregation equals the sum over all units of the units' deviations from the exactly even distribution. Using the formulation in Carrington and Troske (1997):

$$D = \frac{1}{2} \sum_w \left| \frac{s^w - m^w}{N(1-p)} - \frac{m^w}{Np} \right| = \sum_w |D^w|$$

where D^w is the deviation from evenness in unit w . If we denote the probability that unit w has exactly n minorities by $p^w(m=n)$ we get

$$E(D) = \sum_w \sum_{n=0}^{s^w} p^w(m^w = n) \frac{1}{2} \left| \frac{s^w - n}{N(1-p)} - \frac{n}{Np} \right|$$

We can calculate an expected value for the Duncan-index in the absence of segregation if we group units by X , with a specific probability $p(x)$ for each x that an individual belonging to the unit is a minority. It is required that $p(x)$ does not vary within a unit. Empirically, $p(x)$ is given by the realized fraction of minorities in group x :

$$p(x) = \frac{\sum_{w \in x} m^w}{\sum_{w \in x} s^w}$$

The expectation is consequently generated by using the binomial distribution and summing over all units over all possible realizations:

$$E(D) = \sum_w E\left(|D^w(x^w)|\right) = \sum_w \sum_{n=0}^{s^w} Bin(n, s, p(x^w)) \frac{1}{2} \left| \frac{s^w - n}{N(1-p)} - \frac{n}{Np} \right|$$

Definitions of the Duncan index and the Gini coefficient

Massey & Denton (1988) give the following definitions of the Duncan index (D) and the Gini coefficient (G).

$$D = \sum_{i=1}^n \left[\frac{t_i |p_i - P|}{2TP(1-P)} \right]$$

$$G = \sum_{i=1}^n \sum_{j=1}^n \left[\frac{t_i t_j |p_i - p_j|}{2T^2 P(1-P)} \right]$$

where t_i and p_i are the total population and minority proportion of unit i , T is the total population, and P is the minority proportion in the total population. The total population is distributed over n units.