Interpretation problems with changes in indices based on categorizations

Frijters, P.

2000

TI 2000-031/3
Tinbergen Institute Discussion Paper

Interpretation Problems with Changes in Indices based on Categorizations

Paul Frijters

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Keizersgracht 482
1017 EG Amsterdam
The Netherlands
Tel.: +31.(0)20.5513500
Fax: +31.(0)20.5513555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31.(0)10.4088900
Fax: +31.(0)10.4089031

Interpretation problems with changes in indices based on categorizations.

Paul Frijters*

Department of Economics, Free University, Tinbergen Institute,

Amsterdam, The Netherlands

Abstract

In this paper it is argued that occupational and organizational codes maximize the correspondence between activities and easily observable characteristics at the time of their development. Over time the codes become less relevant, leading to the false impression that the segregation of individuals is declining.

JEL code: J31, C43

* E-mail: pfrijters@econ.vu.nl. Ph: +31-20-4446155. URL: http://www.econ.vu.nl/medewerkers/pfrijters

1 Introduction

In the analysis of segregation and discrimination issues, a lot of use has been made of occupational and organizational codes as proxies for different types of activities. Although there are exceptions, the usual finding for the U.S. and other researched countries seems to be that the amount of segregation of individuals with particular characteristics over these occupations or organizations has gone down since the base year when the codes were devised (e.g. Blau et al., 1998; Crocket, 1996; Jacobsen, 1994; MacPherson and Hirsch, 1999; Melkas and Anker, 1998; Watts, 1995a, 1995b; Weeden, 1998; Grusky and Charles, 1998). In this paper it is argued that these findings do not necessarily imply a decrease in the segregation over the phenomena that these codes were intended to proxy: it is argued that such codes are designed to capture as much as possible real distinctions that exist in the activities of individuals at the time that they were drawn up. As technology progresses, old categorizations then very likely become less relevant and there is a tendency for measures of segregation based on the old categorization to decline over time. The only way out of this would be to directly measure the actual item of interest, be it spatial segregation, compensation differences, or the degree of interaction between individuals with different characteristics.

In the next section, the reasoning is captured in a simple model. Conclusions are drawn in the last section.

2 The model

Suppose there are N_A individuals of type A and N_B of type B, where type denotes something of interest to the researcher, such as gender or ethnicity. Suppose that the actual interest of the researcher is the division of individuals of different types into the M possible different activities. As an imperfect tool of assessig the activity of an individual the researcher uses the K$\gg M$ distinct job-descriptions. Denote the proportion of individuals of type A in activity m in period t by $\mathrm{p}_A^{m,t}$ which we assume to remain constant over time. Denote the number of individuals of type A with job-description k in period t by $\mathrm{n}_A^{k,t}$. As to the distribution of individuals over job-descriptions, it is assumed that $\lim_{K,N_A\to\infty} \frac{n_A^{k,t}}{N_A p_A^{m,t}} = \lim_{K,N_B\to\infty} \frac{n_B^{k,t}}{N_B p_B^{m,t}} = 0$ for all k, t, and m. This essentially means we abstract from situations where activities encompass only very few people lumped in a few job-descriptions or where the number of individuals is very small, which would lead to finite number problems of the sort identified by Carrington and Troske (1997).

The researcher is interested in an index y_t that is defined as

$$y_t = y(N_A, N_B, p_A^{1,t} - p_B^{1,t}, .., p_A^{M,t} - p_B^{M,t})$$

where $\frac{\partial y}{\partial |p_A^{m,t} - p_B^{m,t}|} > 0$. This formulation of y encompasses many of the segregation indices often used, such as the Index of Dissimilarity, the Duncan Segregation Index and the Karmel\MacLachan Index. In reality this index will not change.

Activities change over time: new technologies change the actual activities of individuals with prescribed tasks and the division of actual activities over jobs within firms constantly changes. One way of capturing this is to suppose that the activities are associated in each point in time with well-defined, clearly observable job-characteristics (such as titles), but that the mapping of actual activities and these characteristics constantly shifts. This is formalised by the assumption that the combination $(n_A^{k,0}, n_B^{k,0})$ is a 'package' that can switch positions over time: at the end of each period, with a probability δ, the combination $(n_A^{k,0}, n_B^{k,0})$ switches places with any particular package from the set of packages $C=\{(n_A^{1,0}, n_B^{1,0}), .., (n_A^{k-1,0}, n_B^{k-1,0}), (n_A^{k+1,0}, n_B^{k+1,0}), .., (n_A^{K,0}, n_B^{K,0})\}$.

This means that there is a probability of ρ =(1-(1-δ)^K) that ($n_A^{k,0}$,$n_B^{k,0}$) is switched in period 1. If ($n_A^{k,0}$,$n_B^{k,0}$) and ($n_A^{j,0}$,$n_B^{j,0}$) are switched at the end of period 0, then ($n_A^{k,1}$,$n_B^{k,1}$) =($n_A^{j,0}$,$n_B^{j,0}$) and ($n_A^{j,1}$,$n_B^{j,1}$)=($n_A^{k,0}$,$n_B^{k,0}$). The same process holds in all subsequent periods.

Now, denote the set of job-descriptions that corresponds to an activity m in period t as $S_{m,t}$. There holds that the union of these sets span all the possible job-descriptions and there are no overlaps, i.e. $S_{1,t} \cup S_{2,t} \cup .. \cup S_{M,t} = \{1, 2, ..., K\}$ and $S_{m,t} \cap S_{l,t}$ is empty iff m$\neq l$. If ($n_A^{k,0}$,$n_B^{k,0}$) and ($n_A^{j,0}$,$n_B^{j,0}$) are switched at the end of period 0 and k$\in S_{m,0}$ and j$\in S_{l,0}$, then k$\in S_{l,1}$ and j$\in S_{m,1}$. Hence, a sort of random mixing over time is assumed to exist between the actual activities of individuals and their job-description.

Consider now the work of the researcher who wants to identify individuals in different activities on the basis of their job-description. In period 0, good research perfectly informs the researcher about the sets $S_{m,0}$ for all m$\in \{1, .., M\}$. In that period and in further periods the researcher then uses $\hat{p}_A^{m,t} = \frac{1}{N_A} \sum_{i \in S_{m,0}} n_A^{i,t}$ as estimates for $p_A^{m,t}$. An analogous formula holds for $p_B^{m,t}$. In period 0, $\hat{p}_A^{m,0}$ =$p_A^{m,0}$. In period 1, ($n_A^{k,1}$,$n_B^{k,1}$) equals ($n_A^{k,0}$,$n_B^{k,0}$) with probability (1-ρ) and equals a random draw from C with probability ρ. As the number of packages is assumed very large, $n_A^{m,t}$ is the same as $n_A^{m,0}$ with

probability (1-ρ)^t and approximately[1] equals a random draw from all $n_A^{\cdot,t}$ with probability 1-(1-ρ)^t. If we look at the situation where K and $N_\cdot$ become very large, but where $\frac{N}{K}$ does not converge to 0, the distribution of this random draw has a mean equal to $\frac{N_A}{K}$ and an unknown but finite variance σ_A^2. The distribution of $\hat{p}_A^{m,t}$ can then be described as

$$\begin{aligned} \hat{p}_A^{m,t} &= \frac{1}{N_A} \sum_{i \in S_{m,0}} n_A^{i,t} \\ E[\hat{p}_A^{m,t}] &= (1-\rho)^t p_A^{m,0} + [1-(1-\rho)^t] \frac{Z_{S_{m,0}}}{K} \\ Var[\hat{p}_A^{m,t}] &= [1-(1-\rho)^t] \frac{Z_{S_{m,0}}}{N_A^2} \sigma_A^2 \end{aligned}$$

where $Z_{S_{m,0}}$ denotes the number of elements of $S_{m,0}$. From this we may see that the variance of $\hat{p}_A^{m,t}$ is increasing over time (though equal to 0 when K and hence also N_A goes to infinity) and that $\hat{p}_A^{m,t}$ converges to $\frac{Z_{S_{m,0}}}{K}$. Now for $\hat{p}_B^{m,t}$ there holds

[1]The approximation is due to the fact that a package cannot be switched with itself. Because $n_A^{k,1}$ is very small compared to N_A and δ is very small absolutely, this has no consequences.

$$
\begin{aligned}
\hat{p}_B^{m,t} &= \frac{1}{N_B} \sum_{i \in S_{m,0}} n_B^{i,t} \\
E[\hat{p}_B^{m,t}] &= (1-\rho)^t p_A^{m,0} + [1-(1-\rho)^t] \frac{Z_{S_{m,0}}}{K} \\
Var[\hat{p}_B^{m,t}] &= [1-(1-\rho)^t] \frac{Z_{S_{m,0}}}{N_B^2} \sigma_B^2
\end{aligned}
$$

which means $\hat{p}_B^{m,t}$ converges to $\frac{Z_{S_{m,0}}}{K}$ over time. This implies that as long as $|p_A^{m,0} - p_B^{m,0}| > 0$, then $\lim_{K,N_A,N_B \to \infty} \mathrm{P}[|\hat{p}_A^{m,t} - \hat{p}_B^{m,t}| < |\hat{p}_A^{m,s} - \hat{p}_B^{m,s}|] \to 1$ for all m and s<t. Essentially this means that as long as activities are initially spread out over enough job-characteristics, the population is large enough, and there is some segregation in the initial position, that measured differences in proportions of individuals in these activities will decrease over time due to the random switching. Because the same reasoning holds for all the estimated proportions of individuals of different types in different activities, $\lim_{K,N_A,N_B \to \infty} \mathrm{P}[y_t - y_s < 0] \to 1$ iff s<t. Hence, the estimated segregation index will most likely show a decrease over time.

3 Conclusion

This paper makes a single simple point: administrative distinctions between activities, such as occupational or organizational codings, by design most accurately capture real distinctions at the time that they are thought up and are therefore very likely to become less accurate thereafter. Analyses of segregation or discrimination based on these distinctions will then be biased towards the conclusion that segregation and discrimination is reducing over time.

This type of problem may also occur in other analyses that use categorizations. Some indirect indications of the relevance of the mechanism described in this paper comes from the experience with years of education categories as explanatory variables: one of the findings in recent years has been a sharp increase in the within-wage variance of individuals in education-years categories (e.g. Bound and Johnson, 1992). The amount of wage variance explained by education has also fallen. Both findings could be interpreted as evidence that education nowadays captures the actual skill level of individuals less well than it used to, possibly because of an increasing divergence in the quality of education or because of an increase in unmeasured ways in which individuals increase their knowledge. Given the reasoning in this paper, such divergence

in time of the correspondence between a used category and the item that one wishes to capture is an integral part of the attempt to use indirect indicators of the actual items of interest, such as skills or activities. It is a consequence of the fact that good researchers in a changing environment will find the indicators that correspond best with what they are actually interested in at the time of their research.

References

1. Blau, F.D., Simpson, P., Anderson, D. (1998), 'Continuing Progress? Trends in Occupational Segregation in the United States over the 1970s and 1980s', *Feminist Economics* 4(3), 29-71.

2. Bound, J., Johnson, G. (1992), "Changes in the structure of wages in the 1980's: an evaluation of alternative explanation", *American Economic Review*, vol. 83(3), pp. 371-392.

3. Carrington, W.J., Troske, K.R. (1997), 'On Measuring Segregation in

Samples with Small Units', *Journal of Business and Economic Statistics* 15, 402-09.

4. Crockett, G.V. (1996), 'Shifts in Gender Segregation in Professional Occupations in Australia–1981 to 1991', *Australian Bulletin of Labour* 22, 265-74.

5. Grusky, D.B., Charles, M. (1998), 'The Past, Present, and Future of Sex Segregation Methodology', *Demography* 35, 497-504.

6. Jacobsen, J.P. (1994), 'Trends in Work Force Sex Segregation, 1960-1990', *Social Science Quarterly* 75, 204-11.

7. MacPherson, D.A., Hirsch, B.T. (1999), 'Wages and gender composition: why do women's jobs pay less', *Journal of Labor Economics* 13, 426-71.

8. Melkas, H., Anker, R. (1998), *Gender equality and occupational segregation in Nordic labour markets*, ILO, Waldorf.

9. Watts, M.J. (1995a), 'Trends in Occupational Segregation by Race and Gender in the U.S.A., 1983-92: A Multidimensional Approach', *Review of Radical Political Economics* 27, 1-36.

10. Watts, M. (1995b), 'Divergent Trends in Gender Segregation by Occupation in the United States: 1970-92', *Journal of Post Keynesian Economics* 17, 357-79.

11. Weeden, K.A. (1998), 'Revisiting Occupational Sex Segregation in the United States, 1910-1990: Results from a Log-Linear Approach' *Demography* 35, 475-87.