AN ALGORITHM TO REDUCE THE OCCUPATIONAL SPACE IN GENDER SEGREGATION STUDIES

NEUS HERRANZ,[a] RICARDO MORA[b]* AND JAVIER RUIZ-CASTILLO[b]

[a] *Department of Economics, University of Illinois at Urbana-Champaign, USA*
[b] *Departamento de Economıa, Universidad Carlos III de Madrid, Spain*

SUMMARY

This paper presents an algorithm based on the bootstrap to select an admissible aggregation level, that is, the minimum number of occupational categories that yield a gender segregation value not significantly smaller than that obtained from the large number of occupational categories usually available in any data set. The approach is illustrated using labour force survey data for Spain for the comparison of gender segregation in 1977 and 1992, as well as 1994 and 2000. To measure gender segregation, an additively decomposable segregation index based on the entropy concept is used. Despite a substantial simplification in the size of the occupation space, the decrease in the segregation index is very small and not significant, regardless of the year. Consequently, intertemporal changes in gender segregation can be studied using a greatly reduced classification of occupations that permits an easier interpretation of results.

1. INTRODUCTION

Starting from the seminal work by Bergmann (1974), economists have been interested in the problem of occupational segregation by gender, that is, the tendency of women to be segregated into low-pay and low-status occupations. There is no doubt that the extent of gender segregation in the employed population is an important indicator of women's labour market status.[1]

This paper is concerned with the number of occupations one should work with. It is clear that the use of more detailed categories leads to larger index values, since broader categories mask some of the segregation within them (England, 1981). Consequently, researchers have always sought to work with the largest possible occupation space.[2] However, the idea that, *ceteris paribus*, the larger the number of occupations the better, can be questioned from two points of view. First, there is a potential bias due to small cell size (Blau *et al*., 1998): random allocations of individuals across occupations may generate relatively high levels of gender segregation purely by chance. Second, when the number of occupations is very large, results on segregation become difficult to interpret.

The empirical evidence seems to indicate that reducing considerably the number of occupations does not substantially change the results of intertemporal comparisons, nor even the value of a segregation index in a given year. However, most of these studies have reached their conclusions without using statistical criteria.[3] This paper contributes to the literature by presenting two algorithms based on the bootstrap that sequentially aggregate occupations without losing too much information on gender segregation.

The *benchmark algorithm* aggregates, at each step, two occupational categories based solely on their proportion of female workers. Several shortcomings of this algorithm are addressed in what is called the *modified algorithm*. First, occupations are restricted to cluster only within eight major groupings, so that the resulting categories are easy to interpret. Second, the *modified algorithm* is divided into two stages. In the first stage, the large number of occupational categories usually available in any data set are aggregated until the smallest occupation has at least 150 sample observations and it can safely be assumed that the small cell problem has disappeared. In the second stage, an *admissible aggregation level* is selected. The latter is the coarser aggregation level that yields an index—the *core gender segregation*—within bootstrap confidence intervals obtained at the end of the first stage. Third, large sized occupations might unduly influence the aggregation sequence during the second stage of the *modified algorithm*. Finally, as intertemporal comparisons between two years might be sensitive to the list of final occupations of the year taken as reference, the *modified algorithm* takes into account the sum of the square distances between female proportions in each pair of occupations from the years under comparison.

To implement the algorithm, a segregation index based on the family of income inequality indexes introduced by Theil (1971) is used. The relevance of the approach is illustrated with an empirical application using labour force survey data for Spain.

The paper contains four further sections. Section 2 is devoted to the measurement of segregation. The algorithm is described in Section 3, and results of the *modified algorithm* are presented in Section 4. Section 5 offers concluding comments.

2. THE MEASUREMENT OF SEGREGATION

In this section, the index of segregation and its decomposition into a *within* and a *between* term are presented. Consider situations in which people with a given characteristic, say a three-digit occupation, could be grouped in terms of a second characteristic, say a two-digit occupation, but not *vice versa*. Let there be J three-digit occupations, indexed by $j = 1, \dots, J$, classified into I two-digit occupational groups, indexed by G_i, $i = 1, \cdots, I$. Let F_{ij} and T_{ij} be the number of females and people of both genders, respectively, in occupation j within group i. Let $F_i = \sum_{j \in G_i} F_{ij}$ and $T_i = \sum_{j \in G_i} T_{ij}$ be the number of females and people in group i, and let $T = \sum_i T_i$ be the total number of people in the employed population. Let $W = F/T$ be the proportion of females in the population, $W_i = F_i/T_i$ the proportion of females in group i, and $w_{ij} = F_{ij}/T_{ij}$ the proportion of females in occupation j within group i. The population is said to be segregated in occupation j in group i whenever w_{ij} differs from W.

In information theory, $I^{ij} = w_{ij} \log(w_{ij}/W) + (1 - w_{ij}) \log((1 - w_{ij})/(1 - W))$ is known as the expected information of the message that transforms the proportions $(W, (1 - W))$ to a second

[3] See, for instance, Jacobs (1989), Jacobsen (1994) and Blau *et al*. (1998). The exception is Deutsch *et al*. (1994), where bootstrap methods are used to check the sensitivity of various summary indices to errors in the classification of individuals in the various occupations.

set of proportions $(w_{ij}, (1 - w_{ij}))$. The value of this expected information is zero when the two sets of proportions are identical; it takes larger and larger positive values when the two sets are more different. Thus, for example, when the employed population is predominantly male (W small), the presence of an all-female occupation j within group i ($w_{ij} = 1$) implies a large value of I^{ij}. This is intuitively reasonable for a measure of segregation.

The index I^{ij} provides what is called a *direct* measure of gender segregation in occupation j within group i in relation to the entire employed population. The weighted average of the I^{ij}'s, $I^* = \sum_i \sum_{j \in G_i} (T_{ij}/T)I^{ij}$, provides a reasonable overall measure of occupational segregation. This bounded[4] measure of overall gender segregation can be decomposed into a *between-group* and a *within-group* term.

The expected information of the message that transforms $(W, (1 - W))$ into the proportions $(W_i, (1 - W_i))$ is given by $I^i = W_i \log(W_i/W) + (1 - W_i)\log((1 - W_i)/(1 - W))$. The weighted average of the I^i's, $I^{\mathrm{B}} = \sum_i (T_i/T)I^i$, can be interpreted as the *between-group* (direct) gender segregation induced at the two-digit occupational level. On the other hand, the expected information of the message that transforms $(W_i, (1 - W_i))$ into the proportions $(w_{ij}, (1 - w_{ij}))$ is given by $I_{ij} = w_{ij} \log(w_{ij}/W_i) + (1 - w_{ij})\log((1 - w_{ij})/(1 - W_i))$. The segregation within group i as a whole is defined by $I_i = \sum_{j \in G_i}(T_{ij}/T_i)I_{ij}$. Thus, the *within-group* segregation in the partition by two-digit occupational groups can be defined as $I^{\mathrm{W}} = \sum_i(T_i/T)I_i$. As shown in Mora and Ruiz-Castillo (2003), it turns out that

$$I^* = I^{\mathrm{B}} + I^{\mathrm{W}} \tag{1}$$

This decomposition is useful because it permits us to evaluate the impact of aggregation on the measurement of gender segregation.[5]

3. THE CORE GENDER SEGREGATION IN THE SPACE OF OCCUPATIONAL AND INDUSTRIAL CHOICES

3.1. The Data

The data comes from the Spanish EPA (*Encuesta de Presupuestos Familiares*), a labour force survey that investigates the economic activity and other characteristics of every household member over 14 years of age.[6] The time period starts in 1977, the first year for which microeconomic data is available in electronic support. In 1993 and 1994 there are fundamental changes in the National Classification of Occupations (NCO) and in the National Classification of Industries (NCI), making it impossible to compare the 1977 data with the period starting in 1994. Therefore, two periods are distinguished: from 1977 to 1992, and from 1994 to 2000.

[4] The entropy of the distribution characterized by the proportions $(W, (1 - W))$ is defined by $E = W \log(1/W) + (1 - W)\log(1/(1 - W))$. This expression is a measure of the gender mix in the population. It takes its minimum value, equal to 0, when $W = 0$; otherwise, E is positive and reaches its maximum value, equal to log 2, when $W = 1/2$. As shown in Mora and Ruiz-Castillo (2003), I^* can take values in the interval $[0, E]$, and E in turn is normalized in the unit interval by taking all logarithms in base 2.

[5] For an alternative decomposition using the Gini segregation index, see Silber (1989), Deutsch *et al.* (1994), and sections 7.4 and 7.5 of Flückiger and Silber (1999). In the decomposition based on the Gini segregation index, the overall segregation is decomposed into three terms: a between-group term, a within-group term and a third interaction term.

[6] See the Appendix for a brief description of the data.

Because the EPA is a household survey rather than a census, there is a relatively low number of two-digit occupations and industries. Thus, the Appendix is devoted to searching for a combination of the two variables leading to the largest possible initial number of occupational/industrial categories, which will be referred to as occupations. Although comparable procedures were applied to both periods, the changes in the NCO and NCI definitions lead to rather different initial number of occupations: 106 in the first period and 301 in the second. The rest of this section studies how far the dimensionality of the occupational space can be reduced.

3.2. A Sketch of the Benchmark Algorithm

To see how the algorithm works, take 1977 as an example. Denote by I^* the index of gender segregation for the 106 initial occupations, and compute bootstrap confidence intervals for I^*. The value for I^* is 27.79, whilst the bootstrapped average value, the 1% and the 99% bootstrapped lower and upper bounds of I^* are 27.95, 27.19 and 28.71, respectively.[7]

Consider now the following aggregation algorithm. In each step, the occupation with the lowest number of observations is aggregated with the occupation with the closest female proportion. Each step defines a certain aggregation level indexed by $n = 105, 104, \ldots, 1$. The occupations remaining after step n are of two types: initial occupations not affected by the algorithm up to that point, and aggregated occupations consisting of two or more initial occupations. Regardless of their type, the remaining occupations after step n are indexed by G_i, $i = 1, \ldots, n$.

Let $I^{\mathrm{B}}(n)$ be the direct gender segregation induced in the G_i categories, where $i = 1, \ldots, n$. Analogously, let $I^{\mathrm{W}}(n)$ be the *within-group* gender segregation term that captures the gender segregation within the G_i categories consisting of two or more initial occupations. By equation (1), in each step n we have that $I^* = I^{\mathrm{B}}(n) + I^{\mathrm{W}}(n)$. In this context, the term $I^{\mathrm{W}}(n)$ can be viewed as the aggregation error committed when the classification into G_i, $i = 1, \ldots, n$ categories is selected as the occupational space. Of course, $I^{\mathrm{W}}(n)$ is a non-decreasing function of n: the higher the aggregation level selected, the greater the aggregation error.

The algorithm is fully defined after selecting a stopping rule. One possibility is to select the largest n' for which $I^{\mathrm{B}}(n')$ is greater than or equal to the 1% lower bound for I^*, which in 1977 is 27.19. This leads to a value of $n' = 96$, which implies that the final aggregation level would consist of only $106 - 96 = 10$ occupations. The gender segregation index associated with such an aggregation level is 27.45.[8]

As it stands, the algorithm sketched above has four shortcomings having to do with (i) difficulties in the interpretation of certain aggregate categories, (ii) the small cell problem, especially in the second period, (iii) the role of large sized occupations, and (iv) the sensitivity of the evolution in gender segregation to the choice of the reference year. The solution to these problems requires the modification of the algorithm, which is presented in the following subsections.

[7] Bootstrapped values are based on 5000 replications of the empirical distribution with replacement.

[8] Compare this criterion based on the bootstrap with the informal procedure used in Blau *et al.* (1998), where one eliminates in succession all occupations with less than 50 or 100 observations with 1980 US Census data. The number of occupations is reduced from 470 to 305 and 218, respectively. The corresponding gender segregation indexes are $I(470) = 67.68$, $I(305) = 65.73$ and $I(218) = 62.89$.

3.3. The Interpretation of Aggregated Categories

At each step, the *benchmark algorithm* permits two occupations to be merged regardless of their content or substantive nature. For example, in the 12th step in 1977, ‘writers and journalists’, which are professional occupations, are clustered with ‘furriers and leather workers’, which are blue collar occupations. Thus, the advantage of having a small number of occupations is offset by the inconvenience created by an unrestricted mixing process.

To ensure the ease of interpretation at each step, the original occupations for each period are classified into eight major groups.[9] At every step, an occupation can only be aggregated within the major group to which it belongs.

3.4. The Small Cell Problem

There are almost three times more occupations in the second than in the first period. Furthermore, the fraction of the population employed in occupations with less than 100 or 150 observations is much larger in the second period (see Table AI in the Appendix). The gender segregation index for the initial 301 occupations in 1994 is 31.24, while the bootstrapped average index value and the 1% lower bound are 31.76 and 30.93, respectively. Thus, the distance between the 1994 initial gender segregation value and its 1% lower bound is only 0.31 index points, smaller than the distance between that initial value and its bootstrapped average value, which is equal to 0.51 index points.[10] These results indicate that the small cell problem is jeopardizing the usefulness of the bootstrap.

To solve this problem, the algorithm is made to consist of two stages. In the first one, the smallest occupation gets aggregated with that within the major group with the closest female proportion. This stage ends when all cells are already larger than a minimal value taken to be equal to 150 observations. In the second stage, the two occupations within a given major group with the closest female proportions, regardless of their size, are aggregated at each successive step. Table I shows the consequences of applying the first stage of the algorithm to the four years of the study.

Although the first stage of the algorithm takes a considerable number of steps, both the absolute and the relative reduction of the gender segregation index from the initial situation is small in all years (see rows 5 and 6 in Table I). Moreover, close inspection of bootstrapped average values and 1% bootstrapped lower bounds suggests that the 150-observation limit imposed is appropriate to avoid small cell problems in the second stage of the algorithm.

3.5. The Role of Large Sized Occupations

Assume that the distance between the female proportions of two large occupations is slightly smaller than the corresponding distance from two smaller occupations. If the algorithm proceeds unrestricted and the two larger categories are aggregated, then the decrease in gender segregation

[9] The original 106 occupations for the first period are classified into the following major groups: (1) ‘agriculture’; two blue collar groups: (2) ‘operators and labourers’ and (3) ‘precision, craft and repair’; two white collar groups: (4) ‘services’ and (5) ‘technical, sales and administrative staff’; (6) professional; (7) managerial and (8) the armed forces. The 301 occupations of the second period are classified into eight somewhat different major groups. First, there is only one blue collar group. Second, the white collar group ‘technical, sales and administrative staff’ is broken down into two groups: ‘technical and administrative staff’ and ‘personnel facing the public’.

[10] A similar problem can be found for the year 2000.

Table I. Number of occupations and gender segregation index values after the first stage of the algorithm. Results for 1977, 1992, 1994 and 2000

	1977	1992	1994	2000
1. Initial occupations	106	106	301	301
2. Remaining occupations after the first stage	77	74	134	149
3. Index value after the first stage	27.78	27.94	31.21	33.12
4. Average bootstrapped value	27.89	28.06	31.43	33.35
5. 1% lower bound	27.10	27.26	30.63	32.57
6. Absolute drop in the index value	0.01	0.05	0.03	0.02
7. Percentage drop in the index value, in %	0.04	0.18	0.1	0.06
8. Number of occupations with 200 or less observations	6	5	23	24
9. Percentage of the sample population in these occupations, in %	1.5	1.4	7.0	8.9

will be larger than if the two smaller occupations had been selected for aggregation. Of course, at any step one could aggregate those two occupations within the same major group for which the drop in the segregation index is smallest. This procedure would make cell size relevant, as desired, but would also allow the non-linearity of the index to affect the sequence of aggregations, an undesirable feature. Moreover, it is preferable to avoid the polarization of the population in a few large occupations within each major group.

To deal with this issue, the following modification is introduced: at each step in the second stage of the algorithm, two occupations will be aggregated only if they do not represent more than 50% of the population of the major group to which they belong.

3.6. The Choice of Reference Year

As can be seen in the first row of Table I, at the end of the first stage of the algorithm the number—and hence the nature—of the occupations in 1977 are different from those of 1992, and the same is true of the years 1994 and 2000. Of course, the same difficulties appear at the end of the second stage of the algorithm. Thus, intertemporal comparisons are not possible without another modification of the algorithm.

A possible solution is to classify the individuals in 1992 (1977) according to the occupations selected by the algorithm in 1977 (1992). In this case, the gender segregation value in 1992 (1977) would tend to be *lower* than that obtained according to the occupations selected by the algorithm with 1992 (1977) data. Thus the change in gender segregation will be biased downwards (upwards). By way of example, the consequences of taking 1977 or 1992 as the reference years are shown in Table II. In the first case, gender segregation would have decreased by 4.3% during the period, while in the second case it would have increased by 6.9%. Clearly, there is an index number problem of an unacceptable order of magnitude.

As an alternative, the following modification is introduced. Take, as an illustration, the data for the first period. Consider step 1 of the first stage of the algorithm. Assume that occupation j is the smallest one, i.e. assume that T_j is the smallest number in the set $\{T_{jk}:\ j = 1, \ldots, 106;\ k = 77,\ 92\}$. With data from a single year, the algorithm would aggregate occupation j with occupation $j' \neq j$ in the same major group in that year with the closest female proportion, i.e. occupation j' would be the one for which the distance $|W_j - W_{j'}|$ is minimized. In the present context, there are two sets of such distances, one for each year. Thus, for each k, let $d(j, j', k) = W_{jk} - W_{j'k}$.

Table II. Number of occupations and gender segregation index values after the second stage of the algorithm. Results for 1977 and 1992

	Gender segregation according to	
	1977 system	1992 system
Number of occupations	27	29
Index value in 1977	27.14	25.55
Index value in 1992	25.98	27.31
Change in gender segregation		
Absolute change	−1.16	1.76
Relative change, in %	−4.27	6.89

A natural criterion is to choose the occupation j' that minimizes the expression

$$D_1 = [d(j, j', 77)]^2 + [d(j, j', 92)]^2 \quad \text{for all } j' \neq j$$

In step n of the first stage, let i_n be the smallest occupation after the previous step. Occupation i_n is aggregated with the occupation that minimizes $D_n = [d(i_n, i, 77)]^2 + [d(i_n, i, 92)]^2$ for all $i \neq i_n$. The occupations remaining after step n are denoted by G_{ik}, $i = 1, \ldots, n$, $k = 77, 92$.

The first stage ends when all occupations become greater than or equal to a minimum number of observations, say 150. For later reference, the direct gender segregation levels in 1977 and 1992 at this level of aggregation are denoted by I_{77} and I_{92}, respectively, while the corresponding 1% bootstrapped lower bounds are denoted by L_{77} and L_{92}.

In each step of the second stage, say step m, the two occupations selected for aggregation are those in a given major group that minimize $D_m = [d(i, i', 77)]^2 + [d(i, i', 92)]^2$, for all $i \neq i'$, subject to the condition that the sum of the employed people in that pair of occupations does not exceed 50% of the employed people in the major group to which they belong.

Let n_{77} be the final number of occupations according to the criterion that the 1977 gender segregation index at this level of aggregation is greater than or equal to L_{77}. Similarly, let n_{92} be the final number of occupations in 1992. The maximum of the two numbers, say n', becomes what is called the *admissible aggregation level* common to both years. The corresponding gender segregation values $I^{\mathrm{B}}_{77}(n')$ and $I^{\mathrm{B}}_{92}(n')$ constitute the *core gender segregation* in each of the two years of this period. A flow diagram of the modified algorithm is shown in Figure 1.

4. THE RESULTS ON CORE GENDER SEGREGATION USING THE MODIFIED ALGORITHM

The result of applying the *modified algorithm* in the first period is illustrated in Figure 2.[11] The thin line represents the sequence of $I^{\mathrm{B}}_k(n)$, $k = 77, 92$ in the *benchmark algorithm*. The solid line represents the sequence of $I^{\mathrm{B}}(n)$ in the *modified algorithm*, whilst the 1% lower bounds appear as dotted lines. Naturally, the restrictions make the aggregation error at each step at least as large

[11] Results are robust to slight changes in the stopping rule criterion in either stage. For example, using 125 or 175 as the minimum number of observations in the first stage had a very small effect on the estimation of the 1% lower bound and did not lead to a change in the results of the algorithm in the second stage. On the other hand, taking the 5% quantile as the lower bound in the second stage of the algorithm did not change the results regarding the occupation space and the information loss.

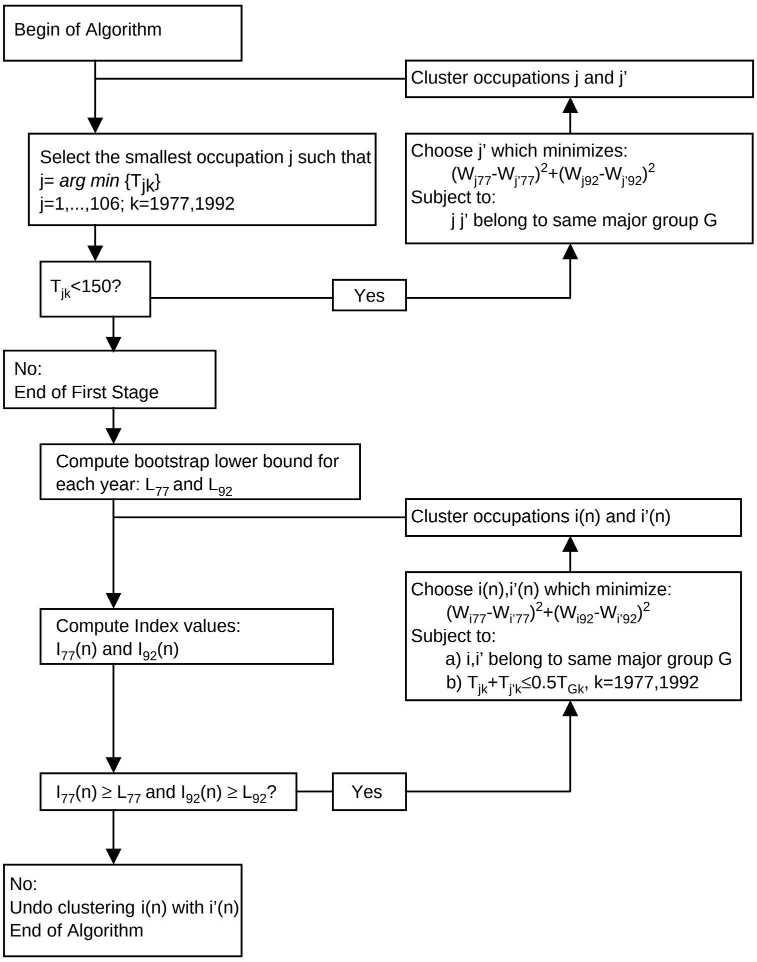

Begin of Algorithm
Cluster occupations j and j'
Select the smallest occupation j such that
j= *arg min* $\{T_{jk}\}$
j=1,...,106; k=1977,1992
Choose j' which minimizes:
$(W_{j77}-W_{j'77})^2+(W_{j92}-W_{j'92})^2$
Subject to:
j j' belong to same major group G
$T_{jk}<150$?
Yes
No:
End of First Stage
Compute bootstrap lower bound for each year: L_{77} and L_{92}
Cluster occupations i(n) and i'(n)
Choose i(n),i'(n) which minimize:
$(W_{i77}-W_{i'77})^2+(W_{i92}-W_{i'92})^2$
Subject to:
a) i,i' belong to same major group G
b) $T_{jk}+T_{j'k}\leq 0.5T_{Gk}$, k=1977,1992
Compute Index values:
$I_{77}(n)$ and $I_{92}(n)$
$I_{77}(n) \geq L_{77}$ and $I_{92}(n) \geq L_{92}$?
Yes
No:
Undo clustering i(n) with i'(n)
End of Algorithm

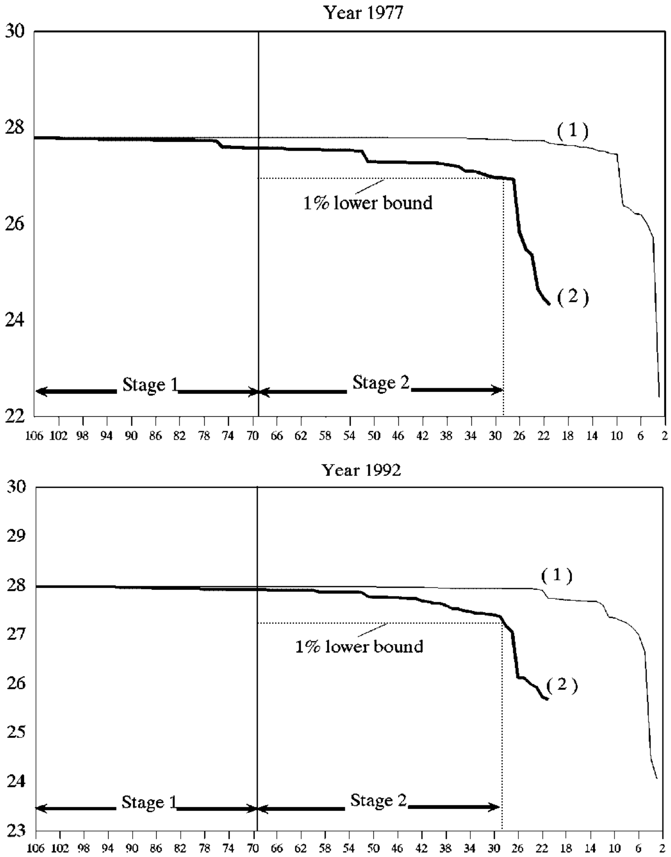
Figure 1. Flow diagram of the modified algorithm for the period 1977 : 1992

as the error in the unrestricted *benchmark algorithm*. However, the aggregation error committed by the *modified algorithm* during the first stage is very small indeed.

The numerical results for both periods are summarized in Table III. After the first stage, the initial 106 and 301 occupations are reduced to 69 and 126, respectively. The gender segregation values at that level of aggregation, as well as their bootstrapped average value and 1% lower bound, are in rows 2 to 4 of Table III. At the end of the second stage, the admissible aggregation level in the first and the second period is reached at 29 and 46 occupations, respectively (see row 7 in Table III).[12]

Once the small cell problem is overcome, the change in gender segregation in the first period is $I_{92} - I_{77} = 0.33$, which represents a slight increase of 1.2% (see rows 5 and 6 in Table III). The

[12] The description of the final categories in terms of the initial occupations in both periods is available upon request. The restrictions imposed on the algorithm ensure that all categories admit a sensible interpretation.



Notes:

(1) The sequence of $I_k^B(n)$, k = 77, 92 for the unrestricted algorithm

(2) The sequence of $I_k^B(n)$, k = 77, 92 for the modified algorithm

Figure 2. The algorithm results for the period 1977 : 1992

difference in core gender segregation in that period is 0.37 (with a 95% confidence interval equal to [−0.65, 1.31]), or an increase of 1.4% (see rows 9 and 10 in Table III). The corresponding magnitudes in the second period are $I_{00} - I_{94} = 1.92$, that represents an increase of 6.2%, and 1.78 (with a 95% confidence interval equal to [0.95, 2.89]), or an increase of 5.8%.[13] Thus, in both periods, the difference in core gender segregation is very close indeed to the change estimated at the end of the first stage.

Finally, the increasing trend in core gender segregation documented in Table III can be accounted for by the interplay of three factors: the increase in the proportion of females in the employed population during both periods, the change in gender segregation in each occupation, and the change in the occupational mix of the economy or the change in the

[13] Confidence intervals were obtained by bootstrapping the original sample but not the algorithm itself. The computational requirements for bootstrapping the entire algorithm in a reasonable length of time are currently too high.

Table III. Number of occupations and gender segregation index values after the first and the second stages of the modified algorithm

First Stage of the Modified Algorithm				
	1977	1992	1994	2000
1. Number of occupations	69	69	126	126
2. Index value	27.58	27.91	31.09	33.00
3. Average bootstrapped value	27.76	28.02	31.30	33.19
4. 1% lower bound	26.95	27.22	30.50	32.39
	First period		Second period	
Change in gender segregation				
5. Absolute change	0.33		1.92	
6. Relative change, in %	1.2		6.2	
Second Stage of the Modified Algorithm				
	1977	1992	1994	2000
7. Number of occupations	29	29	46	46
8. Index value	27.01	27.38	30.65	32.43
	First period		Second period	
Change in gender segregation				
9. Absolute change	0.37		1.78	
10. Relative change, in %	1.4		5.8	

relative demographic importance of each occupation. This analysis is beyond this paper's scope.[14]

5. CONCLUDING COMMENTS

This paper has explored how far it is possible to aggregate an initial list of occupations without reducing the gender segregation value too much. An algorithm has been proposed such that the resulting categories are easy to interpret and, because the final list of occupations is common to the two years under comparison, meaningful intertemporal comparisons can be made. The small cell problem and the role of large size occupations have also been addressed. This technique has been applied using a gender segregation index that is decomposable into a *between* and a *within* term. The within-group term has been identified as the error incurred in each step of the aggregation algorithm.

The empirical application has used labour force survey data for Spain. Two periods are distinguished: from 1977 to 1992, and from 1994 to 2000. After the implementation of the algorithm, the initial 106 and 301 occupations are reduced to 29 and 46 occupations in the first and the second period, respectively. Despite this large simplification in the size of the occupation space, the decrease in the segregation index is very small and not significant.

Finally, it should be noted that the proposed algorithm could be used with any other index of gender segregation. However, the choice of index should be done with care. For example, the

[14] For a study where individual data on occupations during the first period are combined with human capital characteristics, see Mora and Ruiz-Castillo (2003).

most popular index of occupational segregation, Duncan and Duncan's (1955) dissimilarity index, does not change as long as the occupations aggregated at any step are both either female- or male-dominated. Therefore, it is always possible to reduce the number of occupations to be at most twice the number of major groupings. Thus, in this case the algorithm amounts to the choice of the major groupings and loses its appeal.

ACKNOWLEDGEMENTS

We wish to thank two anonymous referees for their helpful comments. Javier Ruiz-Castillo and Ricardo Mora acknowledge financial support from the Spanish Instituto de la Mujer, Project Number 16/97. Ricardo Mora acknowledges financial support from DGI, Grant BEC2000-0170. Neus Herranz acknowledges financial support from the Ministerio de Educación y Ciencia, 2001–2002.

APPENDIX

The EPA is a rotating panel in which each household is interviewed during eight consecutive quarters; thus, one-eighth of the sample is renewed every quarter. In this paper, data from the second quarter is taken as representative of the year as a whole. Due to fundamental methodological changes in the definition of both two-digit occupations and industries, two periods must be distinguished: from 1977 to 1992, and from 1994 to 2000. There are 71,864 and 62,332 individual observations in 1977 and 1992, respectively, which can be classified according to the two-digit NCI of 1974 and the two-digit NCO of 1979. Similarly, the data set contains 57,548 individual observations in 1994 and 66,376 in 2000. These observations are classified according to the two-digit NCI of 1993 and the two-digit NCO of 1994. There is a relatively low number of two-digit occupations and industries: 80 and 64 in the first period, and 65 and 59 in the second period, respectively. This Appendix explores the best way of combining the available information on occupations and industries in order to generate a large list of occupational/industrial categories from which the analysis in the text can proceed.

The simple product of occupations times industries yields $80 \times 64 = 5120$ and $66 \times 59 = 3894$ cells. The fact that only about 18% of cells have more than 25 observations leads us to expect that the gender segregation index defined in this largest possible space is subject to large bias due to small cell size. Taking the year 1977 as an example, the index of gender segregation in this space is 31.87, while the bootstrapped average index value from 1000 empirical sample replications with replacement is 32.61, a considerably higher value; the bootstrapped 1% lower bound is 31.91, a value also greater than 31.87. Thus, the small cell size problem is jeopardizing the usefulness of the bootstrap and another way of combining the information provided by the two variables must be sought.

To assess which variable provides the best basis for a new combination, it is investigated which one has the greatest explanatory value. In 1977, for example, the direct indexes of gender segregation by occupations or industries, computed according to equation (3), are 25.99 and 18.99, respectively. Similar results are obtained for the three remaining years. Consequently, the decision is to take two-digit occupations as the basic partition and combine them with two-digit industries as follows: a given occupation is split into different industries when the set of individuals in each resulting category reaches a certain minimum size; otherwise, the original two-digit occupation

Table A1. The distribution of employed individuals by occupation. Sample statistics for different years

	1977	1992	1994	2000
Number of initial occupations	106	106	301	301
Minimum number of observations	16	12	38	25
1. Occupations with 25 or less observations	2	4	0	1
Percentage over the total, in %	1.9	3.8	0	0.03
2. Occupations with 50 or less observations	13	11	24	26
Percentage over the total, in %	12.3	10.4	8.0	8.6
3. Occupations with 100 or less observations	27	24	163	136
Percentage over the total, in %	25	23	54.2	45.2
4. Occupations with 150 or less observations	33	34	224	196
Percentage over the total, in %	31.1	32.1	74.4	65.1

is left untouched. To simplify the exposition, the resulting occupational/industrial categories will be referred to as ‘occupations’. For comparability reasons, the minimum size should be similar in both periods. To generate a large number of occupations, the minimum size is chosen to be small: 40 observations in the first period and 38 in the second one.

Given the differences in definitions between the two periods, these choices yield 106 occupations in the first period and 301 in the second one. As shown in Table A1, the distribution of the employed population across occupations in the two periods is also very different. The large percentage of individuals in relatively small sized occupations with less than 100 or 150 observations in 1994 and 2000 indicates that the set of occupations in the second period might still suffer from a small cell problem—a question that will be discussed further in the text.

REFERENCES

Albelda R. 1986. Occupational segregation by race and gender, 1958–1981. *Industrial and Labor Relations Review* **39**: 404–411.

Beller A. 1985. Changes in the sex composition of U.S. occupations, 1960–1981. *Journal of Human Resources* **20**: 235–250.

Bergmann B. 1974. Occupational segregation, wages and profits when employers discriminate by race or sex. *Eastern Economic Journal* **1**: 103–110.

Blau F. 1977. *Equal Pay in the Office*. Heath: Lexington, MA.

Blau F, Hendricks W. 1979. Occupational segregation by sex: trends and prospects. *Journal of Human Resources* **12**: 197–210.

Blau F, Simson P, Anderson D. 1998. Continuing progress? Trends in occupational segregation over the 1970s and 1980s. *Feminist Economics* **4**: 29–71.

Deutsch J, Flückiger Y, Silber J. 1994. Measuring occupational segregation. *Journal of Econometrics* **61**: 133–146.

Duncan O, Duncan B. 1955. A methodological analysis of segregation indexes. *American Sociological Review* **20**: 210–217.

England P. 1981. Assessing trends in occupational sex segregation, 1900–1976. In *Sociological Perspectives on Labor Markets*, Berg I (ed.). Academic Press: New York.

Flückiger Y, Silber J. 1999. *The Measurement of Segregation in the Labor Force*. Physica-Verlag: Heidelberg.

Jacobs J. 1989. Long-term trends in occupational segregation by sex. *American Journal of Sociology* **95**: 160–173.

Jacobsen J. 1994. Trends in work force sex segregation, 1960–1990. *Social Science Quarterly* **75**: 204–211.

Mora R, Ruiz-Castillo J. 2003. Additively decomposable segregation indexes. The case of gender segregation by occupations and human capital levels in Spain. *Journal of Economic Inequality* **1**: 147–179.

Silber J. 1989. Factor components, population subgroups and the computation of the Gini index of inequality. *Review of Economics and Statistics* **LXXI**: 107–115.

Theil H. 1971. *Principles of Econometrics*. John Wiley & Sons: New York.

Williams G. 1979. The changing U.S. labor force and occupational differentiation by sex. *Demography* **16**: 73–88.