The Stata Journal (2022)
22, Number 3, pp. 521–556

DOI: 10.1177/1536867X221124471

COMPUTING DECOMPOSABLE MULTIGROUP INDICES OF SEGREGATION

Daniel Guinea-Martin[1] Ricardo Mora[2]

Stata Spanish Meeting
Madrid, October 2022

[1]Department of Sociology, Universidad de Málaga
[2]Department of Economics, Universidad Carlos III Madrid

Outline

1. 2 notions, 4 properties, and 8 indices
2. The data
3. The `dseg` command in Stata

Two notions of segregation, four additive decomposability properties, and eight segregation indices

Two notions of segregation

- Most indices of segregation measure the extent of differences between the proportion of groups (races, genders,...) and the same proportions within each organizational unit (schools, ,occupations, ...): P_{group} vs $P_{\text{group|unit}}$
    - To what extent does the group mixture in the units diverge from the group composition of the population under study?
    - We label this $P_{\text{group|unit}}$: 'group segregation in units', e.g., 'race segregation in schools'.
- Other indices measure how the marginal distribution of units differs from the same distribution within each group: P_{unit} vs. $P_{\text{unit|group}}$: 'unit segregation by group', as in 'school segregation by race'.

Four additive decomposability properties

Unit Decomposability

- Example: in the context of a partition of N schools into K school districts:
- $\Psi^N = \Psi^K + \sum_{k=1}^K \omega_k \times \Psi^{N_k}(k)$
- If $\sum_{k=1}^K \omega_k \begin{cases} = 1: & SUD \\ \neq 1: & WUD \end{cases}$

Group decomposability

- Example: in the context of a partition of G races into $L = 2$ supergroups (whites vs. minorities):
- $\Psi^G = \Psi^L + \sum_{l=1}^L \omega_l \times \Psi^{G_l}(l)$
- If $\sum_{l=1}^L \omega_l \begin{cases} = 1: & SGD \\ \neq 1: & WGD \end{cases}$

Eight decomposable multigroup indices of segregation

	M	NM	Theil's H	$H_{\text{group}\mid\text{unit}}$	Atkinson	$A_{\text{unit}\mid\text{group}}$	Relative	$R_{\text{group}\mid\text{unit}}$
Original citation	Theil & Finizza (1971)	Mora & Ruiz-Castillo (2011)	Theil & Finizza (1971)	Mora & Ruiz-Castillo (2011)	Frankel & Volij (2011)	Here	Carlson (1992)	Here
Notions	Both	Both	$P_{\text{group}\mid\text{unit}}$	$P_{\text{unit}\mid\text{group}}$	$P_{\text{unit}\mid\text{group}}$	$P_{\text{group}\mid\text{group}}$	$P_{\text{group}\mid\text{unit}}$	$P_{\text{unit}\mid\text{group}}$
Properties	SUD, SGD	SUD if $G \leq N$ SGD if $N \leq G$	WUD	WGD	WUD	WGD	WUD	WGD
Commands	`dseg`, `dicseg`†	`dseg`	`dseg`, `seg`, `dicseg`†	`dseg`	`dseg`, `hutchens`†	`dseg`	`dseg`, `seg`	`dseg`

†: Only for the $G = 2$ case.
See Theil and Finizza [1971], Carlson [1992], Reardon and Firebaugh [2002] , Frankel and Volij [2011], and Mora and Ruiz-Castillo [2011] for more details.

The data: A census of the U.S. student enrollment body in public goods

- We use data from the 2017 Common Core of Data (CCD) Local Education Agency Universe Survey
  - Publicly available from the National Center for Education Statistics (NCES)

- All 2017 93,443 public schools, with 45,277,593 students in 16,768 school districts and 51 states
  - Aggregated by sex, race, grade, and school.
  - `student_count` contains the count of students in each cell. In our analyses, we leave aside sex and grades.

```
. use CCD2017_SJdseg.dta
. tabulate race_ethnicity [fweight=student_count], sort missing
```

Race or Ethnicity	Freq.	Percent	Cum.
White	21,675,558	47.87	47.87
Hispanic/Latino	12,059,119	26.63	74.51
Black or African American	6,856,017	15.14	89.65
Asian	2,366,659	5.23	94.88
Two or more races	1,710,347	3.78	98.65
American Indian/Alaska Native	442,643	0.98	99.63
Native Hawaiian/Other Pacific Islander	167,250	0.37	100.00
Total	45,277,593	100.00	

The `dseg` command in Stata

Basic usage

- The simplest call to `dseg` specifies **an index name** and **a notion of segregation**.

- For example, if we have individual-level data where each row is a student ($n = 45,277,593$), we can ask for the standard Theil's H (which is a $P_{\text{group|unit}}$) to measure **race segregation in schools** (with string variables `race_ethnicity` and `schid`

```
. dseg theil race_ethnicity, given(schid)

  Decomposable Multigroup Segregation Indexes

  Differences in race_ethnicity given schid
    Index: Theil´s H

           H
     0.3505
```

- We signal the $P_{\text{group|unit}}$ Theil's H by stating the units (`schid`) in the `given()` option and `race_etnithicy` in the main *varlist* after the `theil` subcommand.

- We can use the `addindex` option to compute in one call the other four indices that follow the same notion of segregation $P_{\text{group|unit}}$.

```
. dseg theil race, given(school) addindex(mutual n_mutual diversity
atkinson) format(%7.6f) fast

   Decomposable Multigroup Segregation Indexes

   Differences in race given school
     Indexes:
     Theil´s H, Mutual Information, Normalized Mutual Information,
     Relative Diversity, Symmetric Atkinson

           H          M          NM          R          A
    0.350479   0.467817   0.240410   0.351159   1.000000
```

  - Option `fast` requires to have contributed command `ftools` installed. Group and unit variables must be numeric.

- The value of H implies that M is 35% the entropy of race.
- The value of NM simply indicates that M is ${}^{0.4678}/_{\log(7)} \times 100 = 24\%$ of its maximum.
- $A_{\text{group|unit}} = 1$ because whenever a race group is absent from one school, that group contributes with its maximum (${}^{1}/_{G}$) to segregation. As no race is present in every single school, the index reaches its maximum value of 1.

- Instead, we write `school, given(race)`, to compute the indices that follow the $P_{\text{school|race}}$ notion of segregation:

```
. dseg theil school [fw=student_count], given(race) addindex(mutual
n_mutual diversity atkinson) format(%9.6f)

   Decomposable Multigroup Segregation Indexes

   Differences in school given race
     Indexes:
     Theil´s H, Mutual Information, Normalized Mutual Information,
     Relative Diversity, Symmetric Atkinson

           H           M          NM           R           A
    0.042076    0.467817    0.240410    0.000024    0.735506
```

- Note the use of the frequency weights with aggregated data (and that the fast option is no longer necessary).
- These are measures of school segregation by race. With the exception of the M and NM their values differ from the measures of race segregation in schools (the results shown before).
- The large value of the Atkinson index reflects that the seven racial categories are present simultaneously in only 22.83% of U.S. schools.
- $H_{\text{unit|group}}$ and $R_{\text{unit|group}}$ are lower than $H_{\text{group|unit}}$ and $R_{\text{group|unit}}$ because they are normalized by the entropy and diversity functions for schools.

Intermediate usage

- Local segregation policies can achieve little because they are capped by the upper bound set by race segregation in school districts.

```
. dseg mutual race [fw=student_count], given(school) addindex(n_mutual theil) within(district)
  Decomposable Multigroup Segregation Indexes
  Differences in race given school
    Indexes:
    Mutual Information, Normalized Mutual Information, Theil´s H
  Between/Within district decomposition
        M      M_B      M_W       NM     NM_B     NM_W        H      H_B      H_W
   0.4678   0.3836   0.0842   0.2404   0.1971   0.0433   0.3505   0.2874   0.0631
```

- As fractions of the overall index, the between and within components are equivalent because NM and Theil's H are normalizations of M:

$$^{38.36}/_{46.78} = {}^{19.71}/_{24.04} = {}^{28.74}/_{35.05} = 0.82$$
$$^{8.42}/_{46.78} = {}^{4.33}/_{24.04} = {}^{6.31}/_{35.05} = 0.18$$

- Only 18% of the value produced by the naive measurement of school racial segregation can be unambiguously attributed to the racial segregation in schools.

- We can also obtain the decomposition for the $P_{\text{unit}|\text{group}}$ indices that are unit decomposable: M and Atkinson.

```
. dseg mutual race [fw=student_count], given(school)
addindex(alt_atkinson) within(district)

   Decomposable Multigroup Segregation Indexes

   Differences in race given school
     Index: Mutual Information
   Differences in school given race
     Index: Symmetric Atkinson
   Between/Within district decomposition

          M       M_B       M_W      AltA    AltA_B    AltA_W
     0.4678    0.3836    0.0842    0.7355    0.5006    0.2349
```

- Given that we set the $P_{\text{group}|\text{unit}}$ notion by choosing `race, given(school)`, we need to use `alt_atkinson` in option `addindex()` to obtain the right unit decomposition.

- When the variables defining the clusters and the units can interchange their roles because they have a nonhierarchical relationship, the index can be decomposed in two ways:

1.
```
. dseg theil race [fw=student_count], given(cbsa)
within(state)

   Decomposable Multigroup Segregation Indexes

   Differences in race given cbsa
     Index: Theil´s H
   Between/Within state decomposition

         H       H_B       H_W
    0.1695    0.1128    0.0568
```

2.
```
. dseg theil race [fw=student_count], given(state)
within(cbsa)

   Decomposable Multigroup Segregation Indexes

   Differences in race given state
     Index: Theil´s H
   Between/Within cbsa decomposition

         H       H_B       H_W
    0.1695    0.1574    0.0121
```

- Note that the sum of the net contributions of states and CBSAs does not equal the value of $H_{\text{race}|\text{CBSA}\times\text{state}}$.

- Instead of partitioning schools into school districts, we could partition the seven races in the 2017 CCD into whites and “minority” students

```
. recode race (1/6=1) (7=2), generate(mrg)
. dseg mutual school [fw=student_count], given(race)
within(mrg)
   Decomposable Multigroup Segregation Indexes
   Differences in school given race
     Index: Mutual Information
   Between/Within mrg decomposition
           M       M_B       M_W
      0.4678    0.2372    0.2306
```

- Only about half of school racial segregation ($^{0.2372}/_{0.4678} \approx .5071$) comes down to the segregation of whites from minority students. The other half originates from segregation among the races in the minority category
- Using `dseg mutual race [fw=student_count], given(school) within(mrg)` does not produce the intended result because it creates a unit space made of all the combinations of `school` and `mrg`.

Advanced usage

- We can design a strategy of multiple calls in order to achieve an assortment of results that may deepen the analysis of segregation.
- For example: we may want to carry out a chain unit decomposition: we first partition schools into school districts and then school districts into states.
- Using the Relative Diversity index:

$$R_{\text{race}|\text{school}} = \text{STATE} + \text{DISTRICT} + \text{SCHOOL}$$

we can get it with two calls to `dseg`:

```
. dseg diversity race [fw=student_count], given(district)
within(state)
  (output omitted)
. dseg diversity race [fw=student_count], given(school)
within(district)
  (output omitted)
```

- The two results are:

```
Decomposable Multigroup Segregation Indexes
Differences in race given district
  Index: Relative Diversity
Between/Within state decomposition
      R       R_B       R_W
 0.2882   0.1087   0.1795
Decomposable Multigroup Segregation Indexes
Differences in race given school
  Index: Relative Diversity
Between/Within district decomposition
      R       R_B       R_W
 0.3512   0.2882   0.0630
```

- Hence,

$$R_{\text{race|school}} = \text{STATE} + \text{DISTRICT} + \text{SCHOOL}$$
$$0.3512 = 0.1087 + \quad 0.1795 \quad + 0.0630$$

- In words, the value of school race segregation in states is 0.1087, but it is $(0.1795/0.1087 - 1) \times 100 = 65.13\%$ larger in districts. Finally, once we control for the effect of states and districts, race segregation in schools accounts for $0.0630/0.3512 \times 100 = 17.94$ of the measurement.

- What if we want to control for the differential race shares in states and in school districts and identify the segregation exclusive from minorities?
- For this task, M is the only instrument in the toolbox because it is additively decomposable in partitions of units and groups. It takes three steps to accomplish this goal.
- As before:

$$\begin{aligned} M &= \text{STATE} + \text{DISTRICT} + \text{SCHOOL} \\ &= M^{\text{race}}_{\text{state}} + \sum_t p_{t\bullet} M^{\text{race}}_{\text{district}}(t) + \sum_d p_{d\bullet} M^{\text{race}}_{\text{school}}(d) \end{aligned}$$

- The term $M^{\text{race}}_{\text{school}}(d)$ is group decomposable:

$$M^{\text{race}}_{\text{school}}(d) = M^{\text{minority vs white}}_{\text{school}}(d) + p_{\text{minority}}(d) \times M^{\text{minorities}}_{\text{school}}(d)$$

- We join equations to obtain:

$$M = \text{STATE} + \text{DISTRICT} + \text{MINORITY VS WHITE} + \text{MINORITIES}$$

- We can obtain this complex decomposition with three calls to `dseg` using the `components` suboption in the option `within()`.

```
. dseg mutual race [fw=student_count], given(district) nolist
within(state) saving(Step1,replace)
```
(*output omitted*)

```
. dseg mutual race [fw=student_count], given(school) nolist
prefix(step2) within(district, components) saving(Step2,replace)
```
(*output omitted*)

```
. dseg mutual school [fw=student_count], given(race) nolist
prefix(step3) within(mrg) by(district) clear
```
(*output omitted*)

```
. merge 1:1 district using Step2.dta
```
(*output omitted*)

```
. generate MINORITIES=step2M_w * step3M_W
. generate MINORITY_WHITE=step2M_w * step3M_B

. collapse (sum) MINORITIES MINORITY_WHITE (mean) M=step2M
```
(*output omitted*)

```
. merge 1:1 _n using Step1.dta
```
(*output omitted*)

```
. rename M_B STATE
. rename M_W DISTRICT

. list M STATE DISTRICT MINORITY_WHITE MINORITIES, abbreviate(15)
```

```
. list M STATE DISTRICT MINORITY_WHITE MINORITIES, abbreviate(15)
```

	M	STATE	DISTRICT	MINORITY_WHITE	MINORITIES
1.	0.4678	0.1505	0.2331	.0385773	.0456264

- Racial segregation in states and districts accounts for around $^{(0.1505+0.2331)}/_{0.4678} \times 100 = 80\%$ of race segregation in schools.

- The contribution to school racial segregation of segregation among minorities only, controlling for the segregation that arises between minorities and whites, and for the segregation due to states and districts, is $.0456264$ or $^{.0456}/_{0.4678} \times 100 = 9.75\%$.

- This is more than half of the segregation fueled by the seven race groups: $^{.0456}/_{(0.0386+0.0456)} = 0.54$.

Bootstrapping and simulation

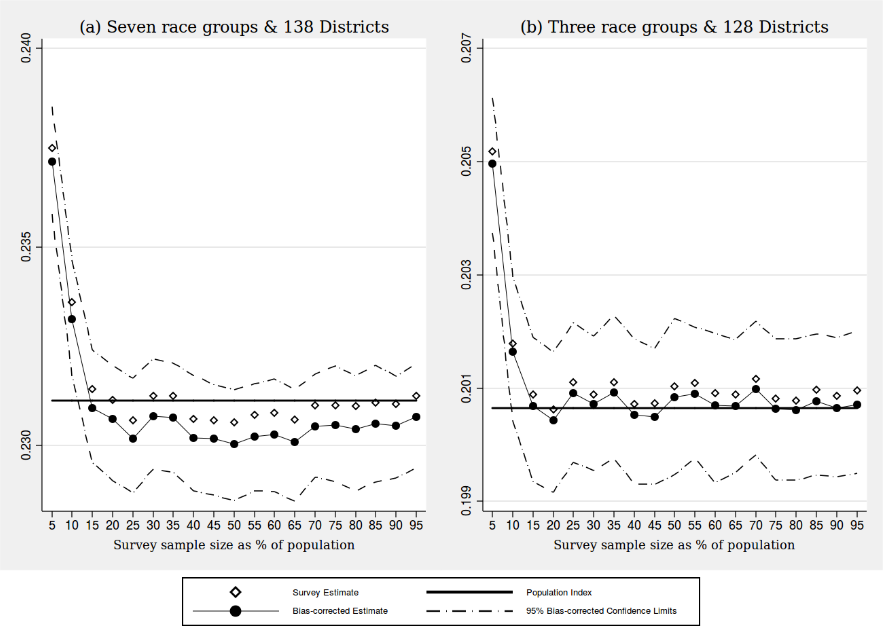
- Survey-based measurements of segregation are finite sample estimates and, therefore, biased and subject to sample variability
- Bootstrap methods can help estimate bias and basic bootstrap confidence intervals for segregation indices
- Option bootstraps() implements the nonparametric bootstrap with individual survey datasets.
- Suppose we have a sample of Alabama schools with sample weights:

```
. expand weights
  (output omitted)

. dseg mutual race, given(district) bootstraps(500) saving("Boots.dta")

  (output omitted)
```

- The new data file `Boots.dta` has 501 observations and includes two variables: (a) `bsn` identifies the bootstrap sample (`bsn==0` refers to the original survey sample); (b) `M` is the index value.
- Data are automatically sorted by `bsn`: `M[1]` corresponds to the M index computed with the original survey sample.


(a) Seven race groups & 138 Districts
0.240
0.235
0.230
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
Survey sample size as % of population
(b) Three race groups & 128 Districts
0.207
0.205
0.203
0.201
0.199
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
Survey sample size as % of population
Survey Estimate
Bias-corrected Estimate
Population Index
95% Bias-corrected Confidence Limits

Randomization tests

- An index computed from a sample can be positive even if the segregation index for the population is zero because of integer constraints (each individual must be uniquely allocated to one unit), and sample variation in small units.

- To discard this possibility, Boisso et al. [1994] propose to use resampling methods (randomization tests) to test that the index is equal to zero.
  - We randomly shuffle the first variable

- Next is an example with 999 replications based on the 10\% Alabamian sample data that we created earlier:

```
. dseg mutual race3, given(district) random(999) clear
  (output omitted)

. generate count=sum(M>=M[1]) in 2/l . generate pvalue=(1+count)/_N
. list pvalue in 2, clean noobs
    pvalue
      .001
```

Thank you!

Dale Boisso, Kathy Hayes, Joseph Hirschberg, and Jacques Silber. Occupational segregation in the multidimensional case: Decomposition and tests of significance. *Journal of Econometrics*, 61(1):161–171, 1994.

Susan M. Carlson. Trends in race/sex occupational inequality: Conceptual and measurement issues. *Social Problems*, 39(3): 268–290, 1992.

David M. Frankel and Oscar Volij. Measuring school segregation. *Journal of Economic Theory*, 146(1):1–38, 2011. ISSN 0022-0531.

Ricardo Mora and Javier Ruiz-Castillo. Entropy-based segregation indices. *Sociological Methodology*, 41(1):159–194, 2011.

Sean Reardon and Glenn Firebaugh. Measures of multigroup segregation. *Sociological Methodology*, 32:33–67, 2002.

Henri Theil and Anthony J. Finizza. A note on the measurement of racial integration of schools by means of informational concepts. *The Journal of Mathematical Sociology*, 1(2):187–193, 1971. doi: 10.1080/0022250X.1971.9989795.