04 – Spatial Demography Concepts and Databases II

J.S. Onésimo Sandoval
Sociology
Saint Louis University

Outline

- Education Attainment Index
- Theil Index of Income Inequality
- GINI Coefficient of Income Inequality
- Normalization of variables and Creating an Index

EDUCATION ATTAINMENT INDEX

Education Attainment

Let's say that 85% of the U.S adult population had at least high school diploma, 27.7% had a bachelor's degree, and 10.2% had a graduate degree.

The education attainment score is .85+.277+.102=1.228.

Note:

Maximum score is 3

Minimum score is 0

You should scale the score for ease of interpretation

Saint Louis City (2006-2010 ACS)

80.6% had at least a high school diploma

26.9% had at least a bachelor's degree

04.0% had at leasta graduate degree.

The education score

is .806+.269+.040=1.115.

STATA Code

1. Educational Attainment for Population 25 Years and Over

Universe: Population 25 Years and Over
Name: T25
Variables:
T025_001: Population 25 Years and Over:
T025_002: Less than High School
T025_003: High School Graduate (Includes Equivalency)
T025_004: Some College
T025_005: Bachelor's Degree
T025_006: Master's Degree
T025_007: Professional School Degree
T025_008: Doctorate Degree

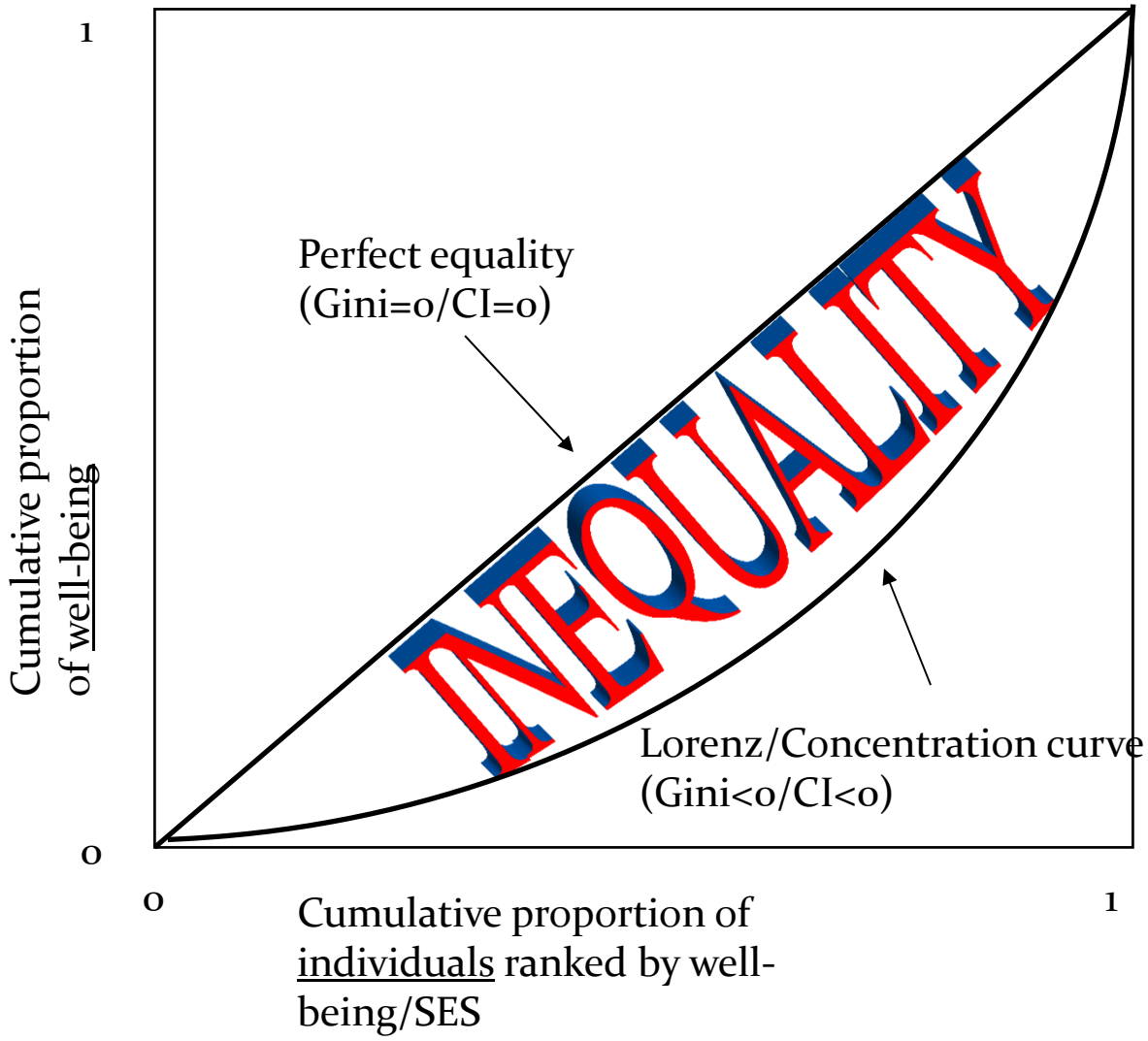
```
*EDUCATION ATTAINMENT.
gen E03=T025_003/T025_001
gen E04=T025_004/T025_001
gen E05=T025_005/T025_001
gen E06=T025_006/T025_001
gen E07=T025_007/T025_001
gen E08=T025_008/T025_001


gen ED_HS=(E03+E04+E05+E06+E07+E08)
gen ED_BS=(E05+E06+E07+E08)
gen ED_GD=(E06+E07+E08)
gen ED_TOT=ED_HS+ED_BS+ED_GD
```

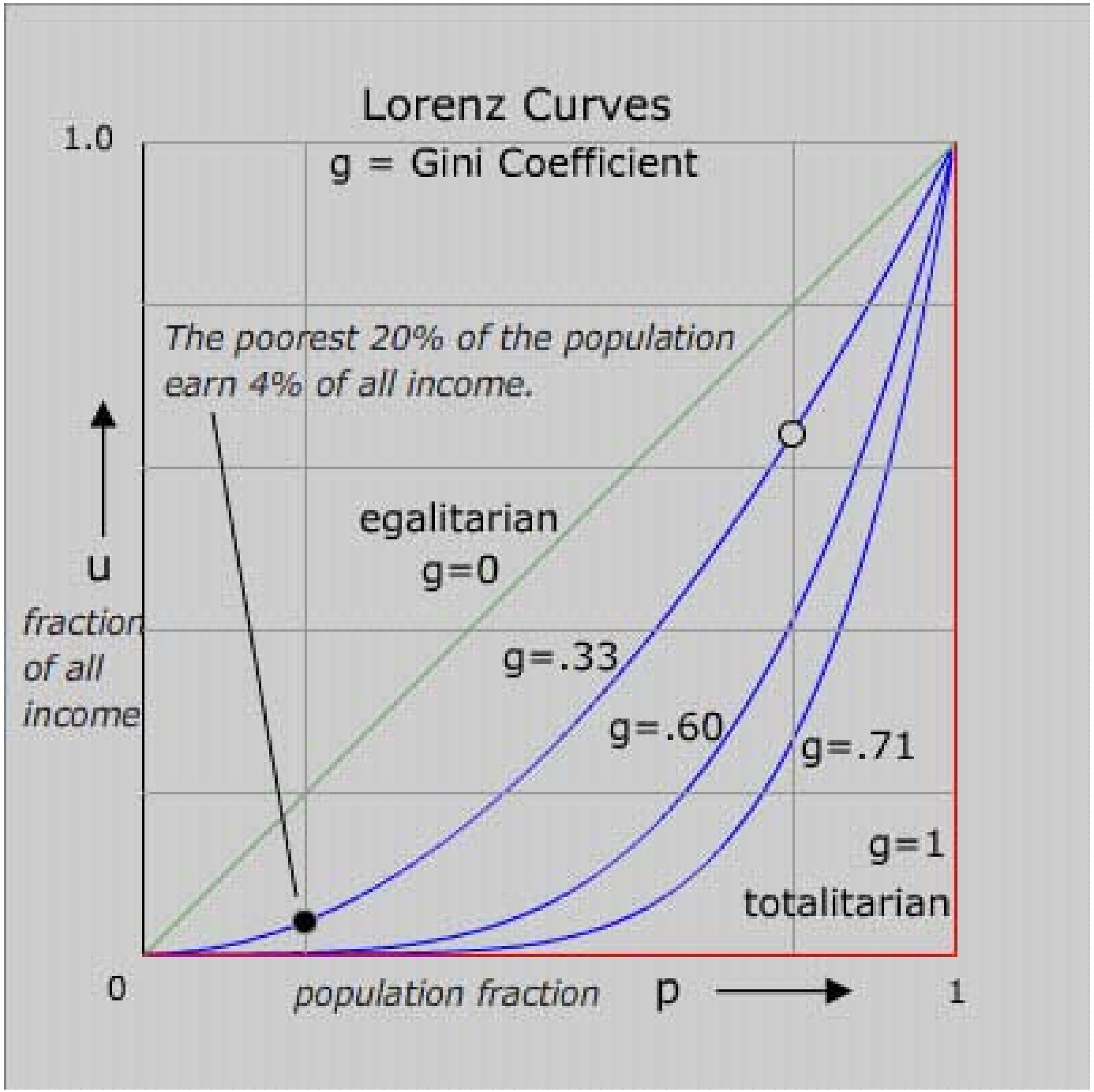
Income Inequality

Lorenz Curve

- Lorenz Curve
  - Max O. Lorenz
    - 1905
      - Represent inequality of the wealth distribution
  - Measures inequalities in the distribution of wealth or income
  - Depict the state of concentration of population and of other demographic aggregates'

Lorenz/Concentration Curve for Well-Being

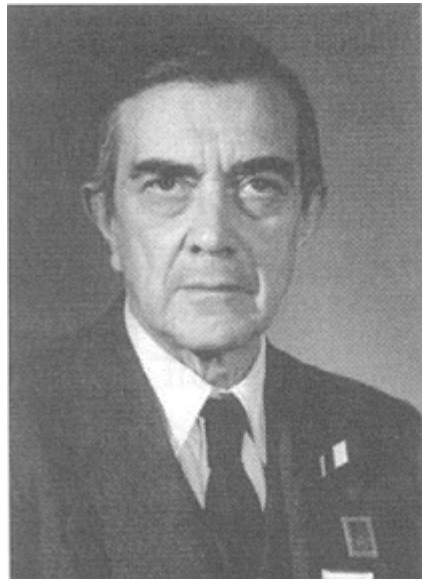


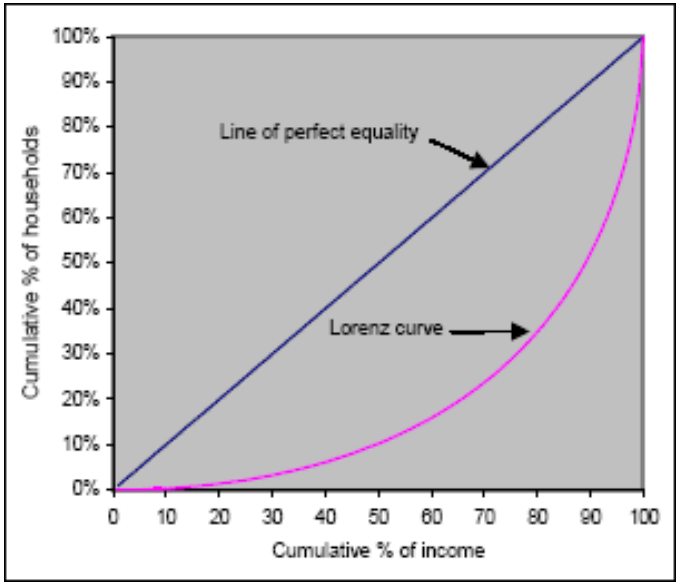

Lorenz Curves
g = Gini Coefficient
1.0
The poorest 20% of the population
earn 4% of all income.
egalitarian
g=0
g=.33
g=.60
g=.71
g=1
totalitarian
u
fraction
of all
income
0
population fraction
p
1

Gini Concentration Ratio

- Corrado Gini – 1912

- Measures the proportion of the total area under the diagonal that lies in the area between the diagonal and the Lorenz Curve.





Gini Concentration Ratio

- The coefficient varies between
- 0 which reflects complete equality
- 1 which indicates complete inequality
- (one person has all the income or consumption, all others have none).
- The Gini coefficient can be used to indicate how the distribution of income has changed within a country over a period of time
- It can be used to compare income distributions across different population sectors as well as countries

GINI Coefficient

$$G_i = 1 - \sum_{i=1}^{N} (X_i - X_{i-1})(Y_i + Y_{i+1})$$

X = Cum. Per. of Income Distribution
(i.e, households or family)

Y = Cum. Per. of Aggregate Income Distribtuion
(need to use midpoint)

Theil-index

- While less commonly used than the Gini coefficient, the Theil-index of inequality has the advantage of being additive across different subgroups or regions in the country.
- The Theil index, however, does not have a straightforward representation and lacks the appealing interpretation of the Gini coefficient.
- The Theil index is part of a larger family of measures referred to as the General Entropy class.

Theil Income Inequality T statistic

$$H_{(y)} = \sum_{i=1}^{m} \left\{ \left(\frac{p_i}{P} \right) * \left(\frac{y_i}{\mu} \right) * \ln \left(\frac{y_i}{\mu} \right) \right\}$$

where :

p_i is the population of the group i,

P is the total population,

y_i is the average income in group i,

μ is the average income across the entire population

Decomposition of the Global City Index

- The sum of the neighborhoods will equal the index for the city
- Between Income Groups (within tracts)
- Within Income Groups (Between tracts)

- For example
- .146 is the index for a city, we may ask how much of the inequality is within the tracts or between the tracts

(1) Calculate the weights for each tract P_{I_i}

(2) Calculate the between group score $A = H_y * P_{I_i} = .062$ (43%)

(3) Calculate the within group score $B = P_{I_i} * \ln\left(P_{I_i} * \left(\sum_{1}^{5} IG / IG_i \right) \right) = .084$ (57%)

Example of Theil Income Inequality

Mid Point Average Income	$25,000	$45,000	$60,000	$75,000	$95,000	Total Population	Average Income	Inequality Within Tract
Neighborhood 1	130	1	1	1	47	180	$43,861	0.213
Neighborhood 2	20	40	60	40	20	180	$60,000	0.054
Neighborhood 3	0	0	0	0	180	180	$95,000	0.000
Neighborhood 4	180	0	0	0	0	180	$25,000	0.000
Neighborhood 5	36	36	36	36	36	180	$60,000	0.085

Example NH 2 - Individuals in the top two income groups contribute positive elements. Individuals in the middle income group contributes nothing to the Theil's T Statistic because the group average salary is equal to the population average. Individuals in the bottom two groups contribute negative elements

STATA Code – Part 1 – Calculate Midpoint for the Income Group

2. Household Income (In <DollarYear> Inflation Adjusted Dollars)

Universe: Households
Name: T56
Variables:

T056_001:	Households:
T056_002:	Less than $10,000
T056_003:	$10,000 to $14,999
T056_004:	$15,000 to $19,999
T056_005:	$20,000 to $24,999
T056_006:	$25,000 to $29,999
T056_007:	$30,000 to $34,999
T056_008:	$35,000 to $39,999
T056_009:	$40,000 to $44,999
T056_010:	$45,000 to $49,999
T056_011:	$50,000 to $59,999
T056_012:	$60,000 to $74,999
T056_013:	$75,000 to $99,999
T056_014:	$100,000 to $124,999
T056_015:	$125,000 to $149,999
T056_016:	$150,000 to $199,999
T056_017:	$200,000 or More

```
*compute midpoint values for each income category aka as Yi.
gen  mp01=5000
gen  mp02=12499
gen  mp03=17499
gen  mp04=22499
gen  mp05=27499
gen  mp06=32499
gen  mp07=37499
gen  mp08=42499
gen  mp09=47499
gen  mp10=54499
gen  mp11=67499
gen  mp12=87499
gen  mp13=112499
gen  mp14=137499
gen  mp15=174999
gen  mp16=325000
```

Data and Methods

- Theil Income Inequality T statistic

$$H_{(y)} = \sum_{i=1}^{m} \left\{ \left(\frac{p_i}{P} \right) * \left(\frac{y_i}{\mu} \right) * \ln \left(\frac{y_i}{\mu} \right) \right\}$$

where:

p_i is the population of the group i,

P is the total population,

y_i is the average income in group i,

μ is the average income across the entire population

STATA Code – Part 2 – Calculate Average Income for census tracts

2. Household Income (In <DollarYear> Inflation Adjusted Dollars)

Universe: Households
Name: T56
Variables:

T056_001: Households:
T056_002: Less than $10,000
T056_003: $10,000 to $14,999
T056_004: $15,000 to $19,999
T056_005: $20,000 to $24,999
T056_006: $25,000 to $29,999
T056_007: $30,000 to $34,999
T056_008: $35,000 to $39,999
T056_009: $40,000 to $44,999
T056_010: $45,000 to $49,999
T056_011: $50,000 to $59,999
T056_012: $60,000 to $74,999
T056_013: $75,000 to $99,999
T056_014: $100,000 to $124,999
T056_015: $125,000 to $149,999
T056_016: $150,000 to $199,999
T056_017: $200,000 or More

```
*Compute total income in each category. This will allow us to calculate average income.
gen I01=T056_002*mp01
gen I02=T056_003*mp02
gen I03=T056_004*mp03
gen I04=T056_005*mp04
gen I05=T056_006*mp05
gen I06=T056_007*mp06
gen I07=T056_008*mp07
gen I08=T056_009*mp08
gen I09=T056_010*mp09
gen I10=T056_011*mp10
gen I11=T056_012*mp11
gen I12=T056_013*mp12
gen I13=T056_014*mp13
gen I14=T056_015*mp14
gen I15=T056_016*mp15
gen I16=T056_017*mp16
gen
INC=(I01+I02+I03+I04+I05+I06+I07+I08+I09+I10+I11+I12
+I13+I14+I15+I16)
gen AVEINC=INC/T056_001
```

Data and Methods

- Theil Income Inequality T statistic

$$H_{(y)} = \sum_{i=1}^{m} \left\{ \left(\frac{p_i}{P} \right) * \left(\frac{y_i}{\mu} \right) * \ln \left(\frac{y_i}{\mu} \right) \right\}$$

where:

p_i is the population of the group i,

P is the total population,

y_i is the average income in group i,

μ is the average income across the entire population

STATA Code – Part 3 – Calculate Part 1 of the formula

```
*Compute Part 1 of the formula - Proportion Breakdown within Neighborhood for Income Groups.
gen P01=T056_002/T056_001
gen P02=T056_003/T056_001
gen P03=T056_004/T056_001
gen P04=T056_005/T056_001
gen P05=T056_006/T056_001
gen P06=T056_007/T056_001
gen P07=T056_008/T056_001
gen P08=T056_009/T056_001
gen P09=T056_010/T056_001
gen P10=T056_011/T056_001
gen P11=T056_012/T056_001
gen P12=T056_013/T056_001
gen P13=T056_014/T056_001
gen P14=T056_015/T056_001
gen P15=T056_016/T056_001
gen P16=T056_017/T056_001
gen PTOT=(P01+P02+P03+P04+P05+P06+P07+P08+P09+P10+P11+P12+P13+P14+P15+P16)
```

Data and Methods

- Theil Income Inequality T statistic

$$H_{(y)} = \sum_{i=1}^{m} \left\{ \left(\frac{p_i}{P} \right) * \left(\frac{y_i}{\mu} \right) * \ln \left(\frac{y_i}{\mu} \right) \right\}$$

where:

p_i is the population of the group i,

P is the total population,

y_i is the average income in group i,

μ is the average income across the entire population

STATA Code – Part 4 – Calculate Part 2 of the formula

```
*Compute Part 2 of the formula  Yi/m m=average income.
gen T101=mp01/AVEINC
gen T102=mp02/AVEINC
gen T103=mp03/AVEINC
gen T104=mp04/AVEINC
gen T105=mp05/AVEINC
gen T106=mp06/AVEINC
gen T107=mp07/AVEINC
gen T108=mp08/AVEINC
gen T109=mp09/AVEINC
gen T110=mp10/AVEINC
gen T111=mp11/AVEINC
gen T112=mp12/AVEINC
gen T113=mp13/AVEINC
gen T114=mp14/AVEINC
gen T115=mp15/AVEINC
gen T116=mp16/AVEINC
```

Data and Methods

- Theil Income Inequality T statistic

$$H_{(y)} = \sum_{i=1}^{m} \left\{ \left(\frac{p_i}{P} \right) * \left(\frac{y_i}{\mu} \right) * \ln \left(\frac{y_i}{\mu} \right) \right\}$$

where:

p_i is the population of the group i,

P is the total population,

y_i is the average income in group i,

μ is the average income across the entire population

STATA Code – Part 5 – Calculate Part 3 of the formula

```
*Compute Part 3 of the fomula ln(Yi/m) m=average income.
gen T201=ln(T101)
gen T202=ln(T102)
gen T203=ln(T103)
gen T204=ln(T104)
gen T205=ln(T105)
gen T206=ln(T106)
gen T207=ln(T107)
gen T208=ln(T108)
gen T209=ln(T109)
gen T210=ln(T110)
gen T211=ln(T111)
gen T212=ln(T112)
gen T213=ln(T113)
gen T214=ln(T114)
gen T215=ln(T115)
gen T216=ln(T116)
```

Data and Methods

- Theil Income Inequality T statistic

$$H_{(y)} = \sum_{i=1}^{m} \left\{ \left(\frac{p_i}{P} \right) * \left(\frac{y_i}{\mu} \right) * \ln \left(\frac{y_i}{\mu} \right) \right\}$$

where:

p_i is the population of the group i,

P is the total population,

y_i is the average income in group i,

μ is the average income across the entire population

STATA Code – Part 6 – Calculate Part 4 of the formula

```
*Compute Part 4 of the formula Part01*Part02*Part03.
gen T301=P01*T101*T201
gen T302=P02*T102*T202
gen T303=P03*T103*T203
gen T304=P04*T104*T204
gen T305=P05*T105*T205
gen T306=P06*T106*T206
gen T307=P07*T107*T207
gen T308=P08*T108*T208
gen T309=P09*T109*T209
gen T310=P10*T110*T210
gen T311=P11*T111*T211
gen T312=P12*T112*T212
gen T313=P13*T113*T213
gen T314=P14*T114*T214
gen T315=P15*T115*T215
gen T316=P16*T116*T216

*Create your income inequality score
gen theil15=(T301+T302+T303+T304+T305+T306+T307+T308+T309+T310+T311+T312+T313+T314+T315+T316)
```

Other STATA Code

3. Poverty Status in 2015 for Children Under 18

Universe: Population Under 18 Years of Age for whom poverty status Is determined

Name: T114

Variables:

T114_001: Population Under 18 Years of Age for Whom Poverty Status Is Determined:

T114_002: Living in Poverty

T114_003: At or Above Poverty Level

4. Poverty Status in 2015 for Population Age 18 to 64

Universe: Population Age 18 to 64 for whom poverty status Is determined

Name: T115

Variables:

T115_001: Population Age 18 to 64 for Whom Poverty Status Is Determined:

T115_002: Living in Poverty

T115_003: At or Above Poverty Level

5. Poverty Status in 2015 for Population Age 65 and Over

Universe: Population Age 65 and Over for whom poverty status Is determined

Name: T116

Variables:

T116_001: Population Age 65 and Over for Whom Poverty Status Is Determined:

T116_002: Living in Poverty

T116_003: At or Above Poverty Level

```
*percent poverty=living in poverty/Population for Whom Poverty Status Is Determined
gen p01=(T114_002+T115_002+T116_002)/(T114_001+T115_001+T116_001)
```

6. Health Insurance

Universe: Civilian Noninstitutionalized Population

Name: T145

Variables:

T145_001: Total:

T145_002: No Health Insurance Coverage

T145_003: With Health Insurance Coverage:

T145_004: Public Health Coverage

T145_005: Private Health Insurance

*percent no health insurance=0 Health Insurance Coverage/Total

gen nhi15=T145_002/T145_001

NORMALIZATION

Index

Conceptual Formula for the Index

(1) $V_i = \left(\frac{X_j - Y_i}{Z_i - Y_i} \right)$

(2) $S_i = \frac{\sum_i^N V_i}{N}$

V_i is the standarized index score for each observation

X_j = actual value for each observation

Y_i =minimum value in the universe of observations

Z_i =maximum value in the universe of observations

S_i is the index sore for each grid cell

N is the number of dimensions

*gen ed_sc15(ED_TOT-min)/(max-min)
*gen mhi_sc15=(log(mhi15)-log(min))/(log(max)-log(min))
*Income Inequality Scale - High values are bad and low values are good
*gen inc_sc15=(theil15-min)/(max-min)
*Recode income Inequality Scale - High values are good and low values are bad
*gen rinc_sc15=1-inc_sc15
*gen index01=(ed_sc15+rinc_sc15+mhi_sc15)/3

STATA Code

```
summarize theil15 ED_TOT mhi15

*    Variable |       Obs       Mean    Std. Dev.       Min        Max
*-------------+--------------------------------------------------------
*     theil15 |       614    .2980129    .0840118   .1302825    .744298
*      ED_TOT |       614    1.315936    .3371135   .5626781   2.433559
*       mhi15 |       614    57267.08    26618.61       9782     196286


*Education Scale - High values are good and low values are bad
gen ed_sc15=(ED_TOT-.5626781)/(2.433559-.5626781)

*Income Inequality Scale - High values are bad and low values are good
gen inc_sc15=(theil15-.1302825)/(.744298-.1302825)

*Recode income Inequality Scale - High values are good and low values are bad
gen rinc_sc15=1-inc_sc15

*Median Income - High is good and low is bad
gen mhi_sc15=(log(mhi15)-log(9782))/(log(196286)-log(9782))


*High is good and low is bad - all variables are equal
gen index01=(ed_sc15+rinc_sc15+mhi_sc15)/3

*High is good and low is bad - education is 50%, inequality is 25% and income is 25%
gen index02=(ed_sc15*.5)+(rinc_sc15*.25)+(mhi_sc15*.25)
```

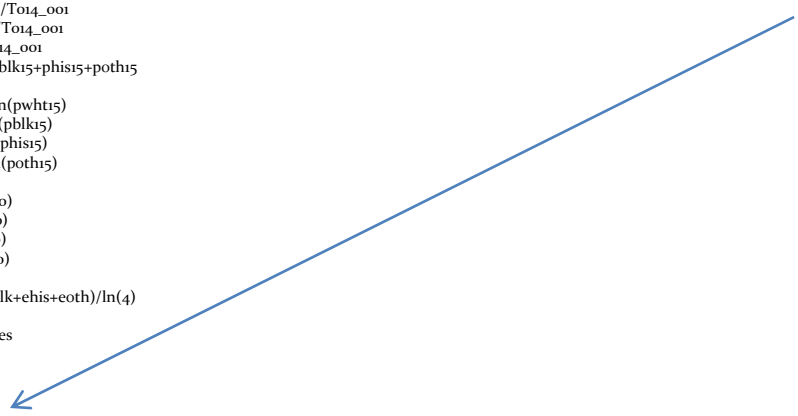
PROCEDURE

Methodology

- Run STATA code to produce part01.xlsx
  - FIPS needs to be renamed GEOID10
- Work to modify shapefiles to the Saint Louis CBSA
  - Create a variable called keep
  - Export this our to a folder
- Run STATA code to read shapefile
- Run STATA code to merge, produce final.xlsx
- Join files in ArcMap

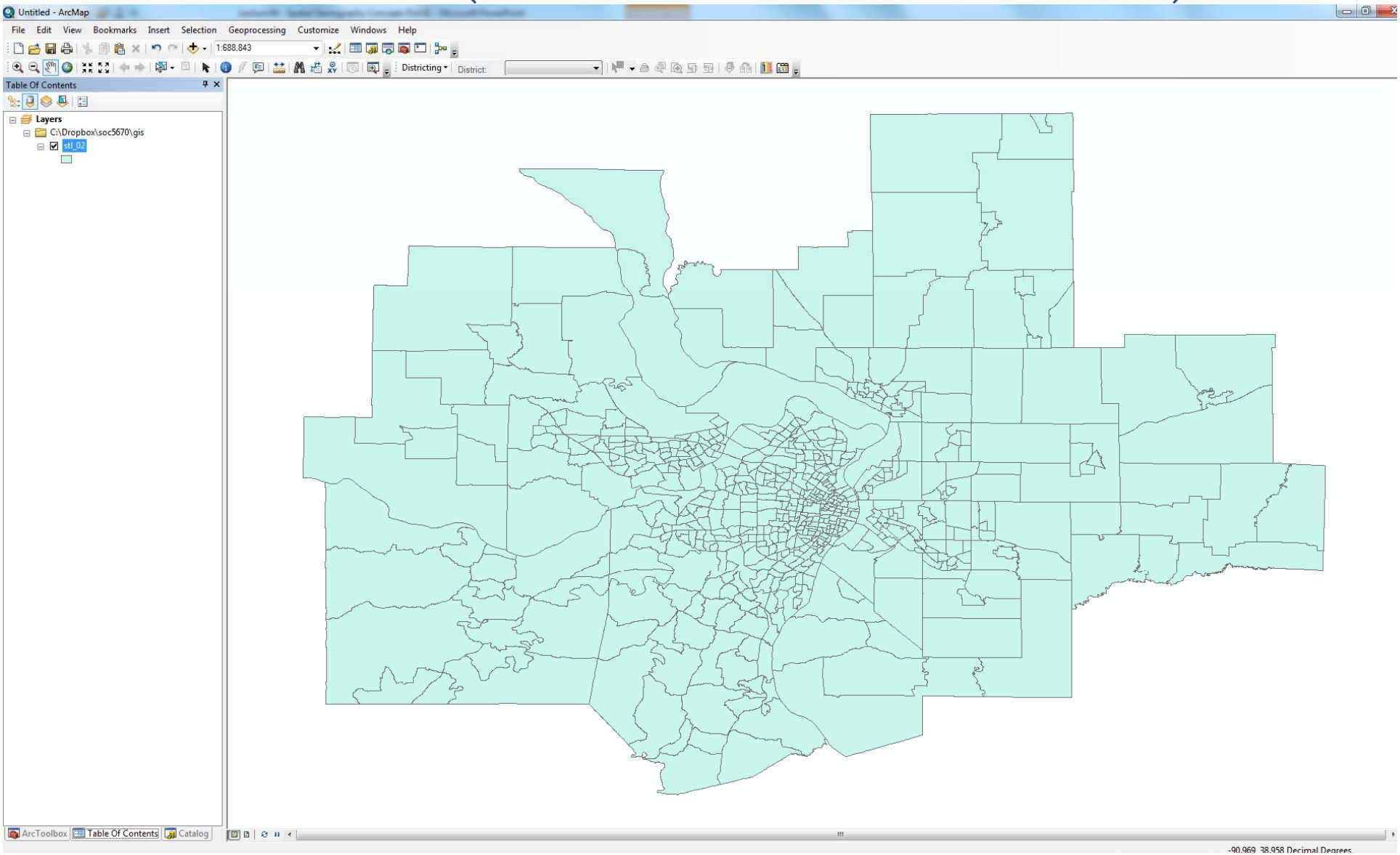
Run STATA code to produce part01.xlsx

```
clear
set more off
infile using "R11318018.dct", using("C:\Dropbox\soc5670\data\R11318018_SL140.txt")
log using analysis.log, replace


*Sex Ratio
gen sr15=T004_002/T004_003


*Dependency Ratio
gen depc15=T008_002+T008_003+T008_004
gen age15=T008_005+T008_006+T008_007+T008_008+T008_009+T008_010+T008_011+T008_012+T008_013+T008_014+T008_015+T008_016+T008_017+T008_018
gen deps15=T008_019+T008_020+T008_021+T008_022+T008_023+T008_024
gen dp15=(depc15+deps15)/age15

gen check=(depc15+age15+deps15)-T008_001

*Child Dependency Ratio
gen cdp15=depc15/age15

*Senior Dependency Ratio
gen sdp15=deps15/age15

*Ageing Index
gen ai15=deps15/depc15


*Racial Diversity Score
gen other15=T014_005+T014_006+T014_007+T014_008+T014_009

gen pwht15=T014_003/T014_001
gen pblk15=T014_004/T014_001
gen phis15=T014_010/T014_001
gen poth15=other/T014_001
gen ptot15=pwht15+pblk15+phis15+poth15

gen ewht15=pwht15*ln(pwht15)
gen eblk15=pblk15*ln(pblk15)
gen ehis15=phis15*ln(phis15)
gen eoth15=poth15*ln(poth15)

recode ewht15(mis = 0)
recode eblk15(mis = 0)
recode ehis15(mis = 0)
recode eoth15(mis = 0)

gen e15=abs(ewht+eblk+ehis+eoth)/ln(4)

*recode some variables
gen wht15=T014_003
gen blk15=T014_004
gen lat15=T014_010
gen GEOID10=FIPS

export excel GEOID10 dp15 cdp15 sdp15 ai15 e15 wht15 blk15 lat15 using "C:\Dropbox\soc5670\data\part01.xlsx", firstrow(variables) nolabel replace,

save "C:\Dropbox\soc5670\data\stl_part01.dta", replace

log close
exit
```



Make sure add the new GEOID10 variable

Work to modify shapefiles to the Saint Louis CBSA (See Previous Lab Notes)



STATA code to read shapefile

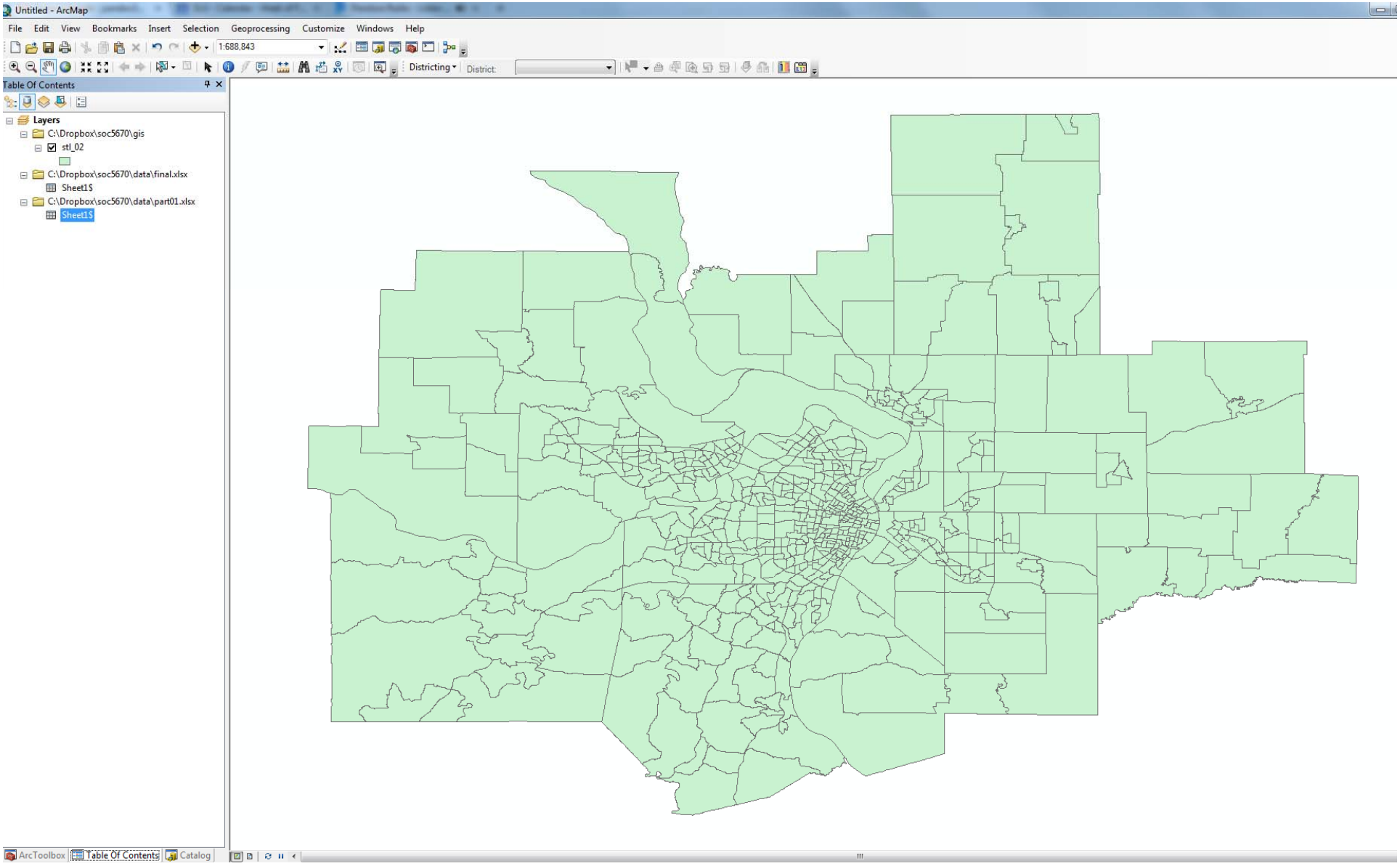
First, we need to export shapefile out the geodatabase
Second, install shp2dta if it is not installed
Third, import shapefile into geodatabase

```
clear
set more off
cd C:\Dropbox\soc5670\gis\
shp2dta using stl_02, database(geo_stl) coordinates(uscd)
```

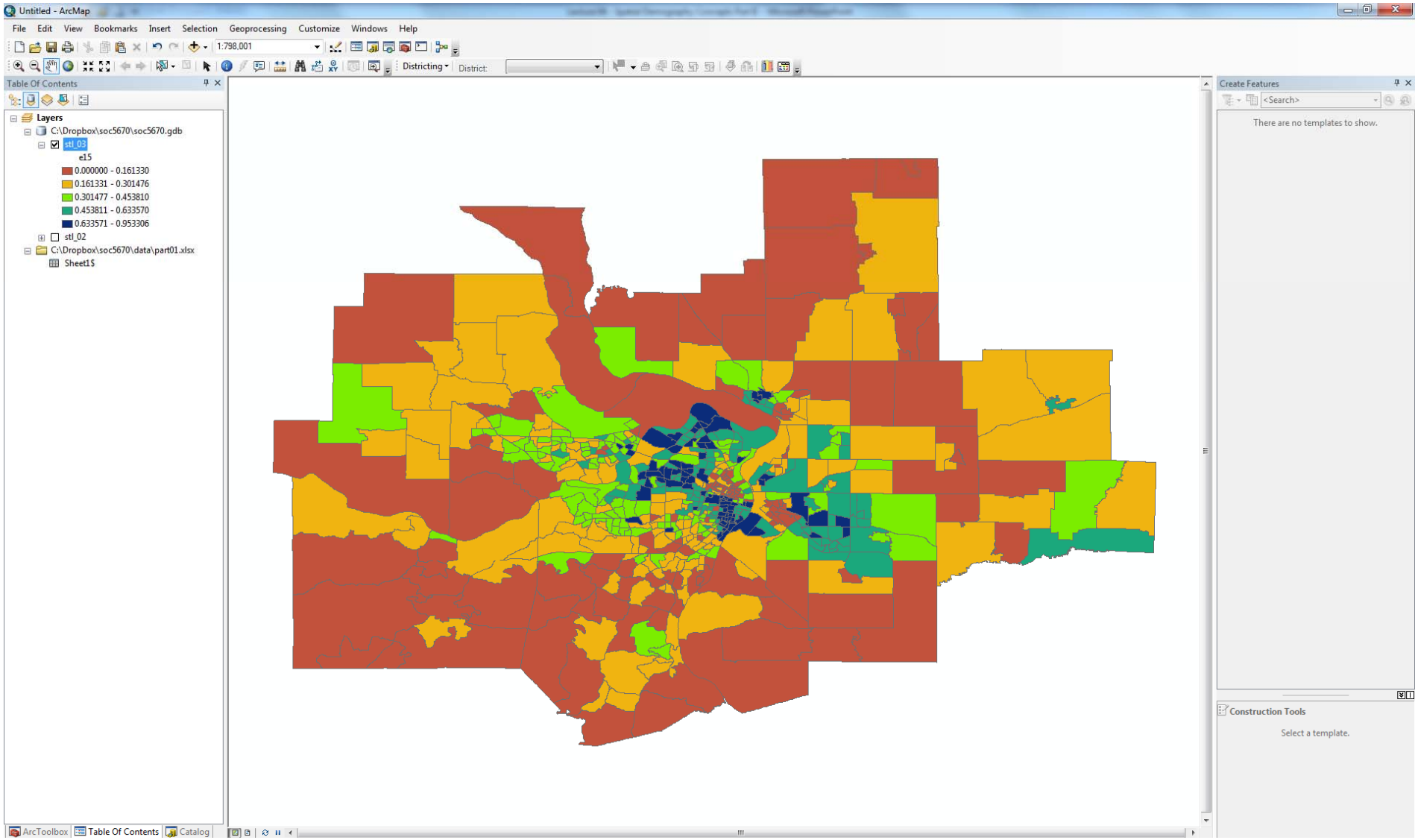
Run STATA code to merge & produce final.xlsx

```
clear
set more off

use "C:\Dropbox\soc5670\gis\geo_stl.dta"

gen GEOID10=GEOID

sort GEOID10

save "C:\Dropbox\soc5670\gis\geo_stl01.dta", replace

use "C:\Dropbox\soc5670\data\part02.dta"

sort GEOID10

merge 1:1 GEOID10 using "C:\Dropbox\soc5670\gis\geo_stl01.dta"

keep if keep==1

*Data Normalization =x-MIN/MAX/MIN replace min and max values.

*gen ED_SCALE=(ED_TOT-min)/(max-.min)
*gen INC_SCALE=(LN(V26)-LN(min))/(LN(max)-LN(min))
*gen INDEX=(ED_SCALE+INC_SCALE)/2

summarize theil15 ED_TOTmhi15


*    Variable |       Obs        Mean    Std. Dev.       Min        Max
*-------------+--------------------------------------------------------
*     theil15 |       614    .2980129    .0840118   .1302825    .744298
*      ED_TOT |       614    1.315936    .3371135   .5626781   2.433559
*       mhi15 |       614    57267.08    26618.61       9782     196286



*Education Scale - High values are good and low values are bad
gen ed_sc15=(ED_TOT-.5626781)/(2.433559-.5626781)

*Income Inequality Scale - High values are bad and low values are good
gen inc_sc15=(theil15-.1302825)/(.744298-.1302825)

*Recode income Inequality Scale - High values are good and low values are bad
gen rinc_sc15=1-inc_sc15

*Median Income - High is good and low is bad
gen mhi_sc15=(log(mhi15)-log(9782))/(log(196286)-log(9782))


*High is good and low is bad - all variables are equal
gen index01=(ed_sc15+rinc_sc15+mhi_sc15)/3

*High is good and low is bad - education is 50%, inequality is 25% and income is 25%
gen index02=(ed_sc15*.5)+(rinc_sc15*.25)+(mhi_sc15*.25)

export excel GEOID10 theil15 ED_TOT p01 nhi15 mhi15 index01 index02  using "C:\Dropbox\soc5670\data\final.xlsx", firstrow(variables) nolabel replace,

save "C:\Dropbox\soc5670\mas_stl.dta", replace
```

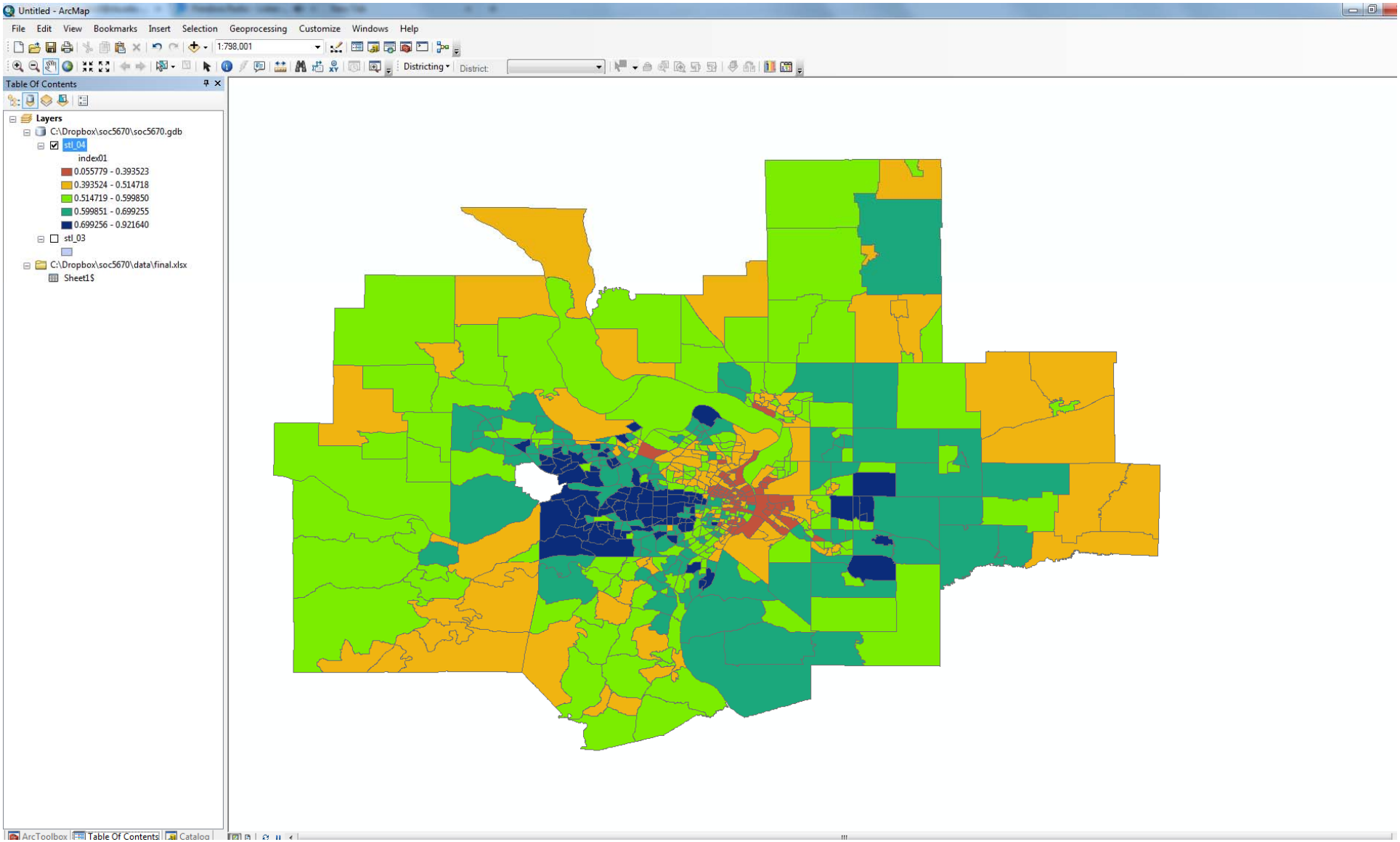
Add shapefile “02,” “final,” and “part01”



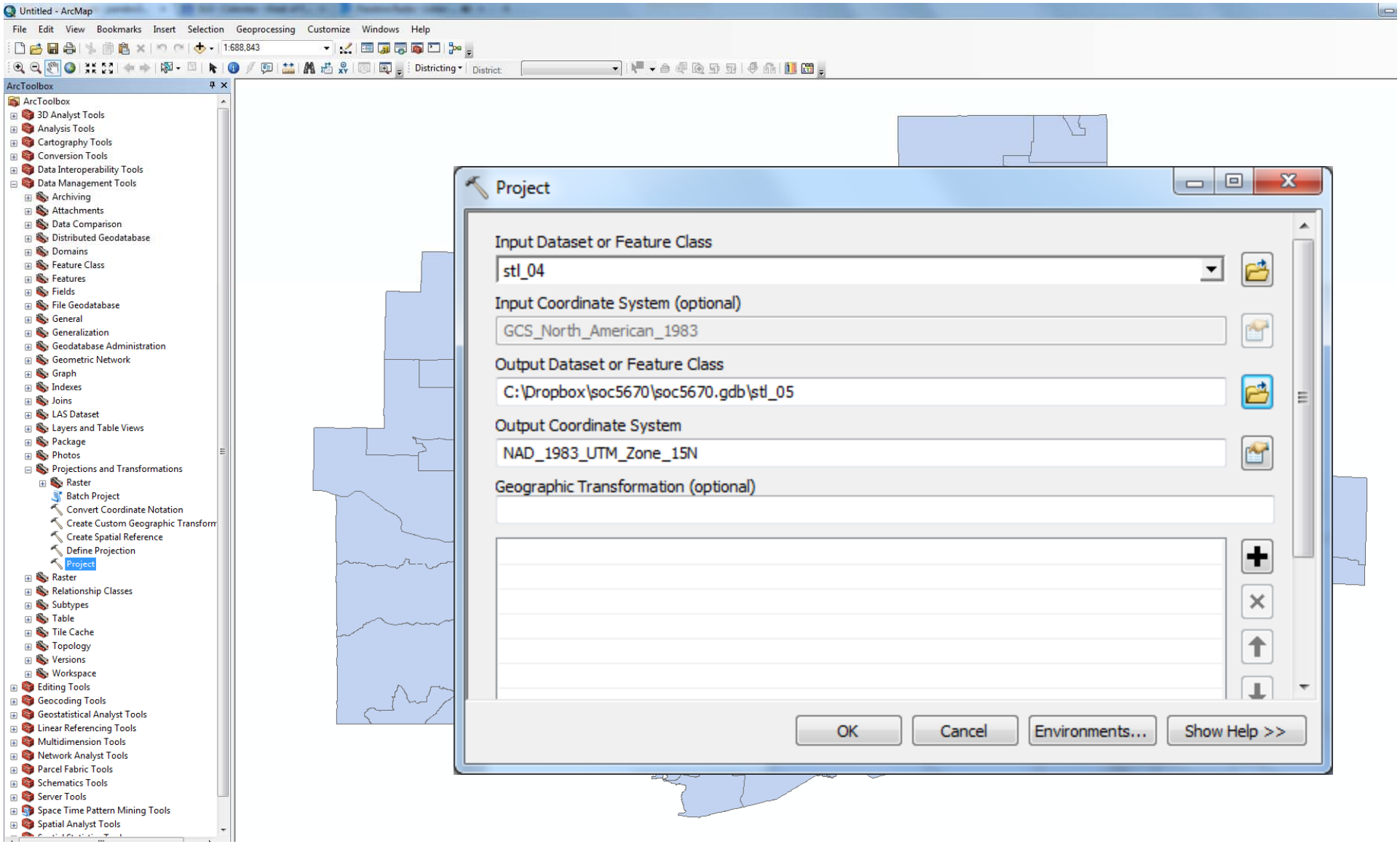
Join "part01" to shapefile "02" and make the join permanent by creating a new shapefile "03"



Join “final” to shapefile “03”and make the join permanent by creating a new shapefile “04”



Project the final version of the shapefile "04"



Change Data Frame Property to match projection