Delaware Valley Regional Planning Commission
Suburbanization of Poverty Study (Title Subject to Change)

Technical Reference

Addison Larson

January 4, 2019

1 Data Sources and Processing

See script `1_clean_data.R` for data sources and processing.

1.1 Data Sources

This study defines "low-income" as persons with incomes below 200% of the Federal Poverty Level (FPL). Because we are interested in spatiotemporal changes in low-income residents, we obtain 1990, 2000, 2010 (midpoint), and 2015 (midpoint) data at the census tract level. The study area includes Burlington, Camden, Gloucester, and Mercer counties in New Jersey; and Bucks, Chester, Delaware, Montgomery, and Philadelphia counties in Pennsylvania.

Data on the number of persons with incomes relative to FPL comes from the 1990 Census Summary Tape File 3 (Table NP121), the 2000 Census Summary File 3a (Table NP088A), and 2012 and 2017 American Community Survey (ACS) 5-year estimates (Table S1701). 1990 and 2000 Census data were obtained from the National Historical Geographic Information System's nominally integrated Time Series Table C20. 2012 5-year estimates were downloaded from American FactFinder, and 2017 5-year estimates were downloaded using the Census API.

1.2 Data Processing

The following values are calculated for all years:

1. Universe of persons for whom poverty status is determined;
2. Counts of persons with incomes up to 200% of FPL and above 200% of FPL and associated margins of error (MOEs) where applicable; and
3. Percentages of persons with incomes up to 200% of FPL and above 200% of FPL and associated MOEs where applicable.

Because census tract boundaries have changed over time, the Longitudinal Tract Database (LTDB) is used to interpolate 1990 and 2000 census tracts to 2010 boundaries. Spatial interpolations of 1990 and 2000 census tracts are approximations of actual percentages of low-income residents, and we have no way of validating these approximations. For this reason, contemporaneous census tracts are used for summary statistics, charts, and maps where possible.

2 Data Analysis

See script `2_clean_data.R` for cluster analysis; and `4_cv.R` for identifying statistically significant change.

2.1 Cluster Analysis, 1990-2015

We are interested in identifying census tracts with marked and noticeable change in low-income residents in recent decades. To this end, census tracts in the study area are divided into five tract typologies based on the 2015 percentage of low-income residents ("baseline") and the percentage change in low-income residents from 1990 to 2015 ("change"). We analyze baseline and change simultaneously because change in itself can be misleading: census tracts with extremely low percentages of low-income residents in 1990 can show large change. For example, a census tract may have an astonishing change of 700%, when the actual shift in low-income residents was from 0.5% in 1990 to 4% in 2015.

Tract typologies are created using model-based clustering along a Gaussian finite mixture distribution; see the `R` package `mclust`. Because the 2015 data comes from the ACS and includes sample error, we run 1,000 cluster analyses with simulated baseline and change values. In each iteration, the baseline and change values vary based upon 2015 estimates and MOEs. Each census tract is assigned a new baseline value, where the mean is the census tract's estimate, the standard deviation is the tract's $\frac{MOE}{1.645}$, and the range of possible new values is normally distributed. The new baseline value is used as an input in creating a new change value.

Table 1 shows the results of the cluster analysis. The 5-group EVV model (ellipsoidal distribution, equal volume, varying shape and orientation) is the clear winner. This result differs from that of simply using the 2015 estimate in calculations, where the optimal model was the 7-group VVI model (diagonal distribution with varying shape and volume), though the VVI model still appears 13 times in the 1,000 iterations of cluster analysis. The results indicate that considering the MOE in computing baseline and change values affects the optimal clustering scheme in a consistent manner.

Identifying changes in census tracts from 1990-2015 is interesting because of the extended timeframe. However, we have no way to verify the validity of 1990 data interpolated to 2010 census tract boundaries. For this reason, we turn to an additional analysis method discussed below.

Table 1: Cluster Analysis Results

Model Name	Number of Groups	n
EVV	5	775
EVV	6	104
EVV	4	76
VVI	5	13
VVI	6	13
VVI	7	13
EVV	7	3
VVE	6	2
VVE	5	1
	TOTAL	1000

2.2 Statistically Significant Increases, 2010-2015

We also test for statistically significant increases in the percentage of low-income residents from 2010 to 2015 at the 95% confidence level, accounting for MOEs. Increases are statistically significant if the below equation exceeds 1.645, which is the one-tailed z-score at a 95% confidence level.

$$\frac{\hat{X}_{2015} - \hat{X}_{2010}}{\sqrt{[SE(\hat{X}_{2015})]^2 + [SE(\hat{X}_{2010})]^2}} \tag{1}$$

See the Census Bureau's Understanding and Using American Community Survey Data, p. 47.

3 References

1. Logan, J. R., Xu, Z., & Stults, B. J. (2014). Interpolating US Decennial Census tract data from as early as 1970 to 2010: A longitudinal tract database. *The Professional Geographer 66*(3), 412-420.

2. Manson, S., Schroeder, J., Van Riper, D., & Ruggles, S. (2018). IPUMS National Historical Geographic Information System (13.0) [Database]. Retrieved from http://doi.org/10.18128/D050.V13.0

3. Scrucca L., Fop M., Murphy T. B., & Raftery A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal 8*(1), 205-233.

4. U.S. Census Bureau. (2018). *Understanding and using American Community Survey data: What all data users need to know.* Washington, DC: U.S. Government Printing Office.