**SSA Technical Information and Methodology**

Statistical Small Areas (SSAs) are clusters of census tracts grouped together based on demographic indicators. SSAs were created for Community Information Now’s “Bexar Data Dive“ platform. This document will explain the methodology behind the creation of SSAs. For a brief overview of the methodology, please visit our [FAQ](https://dive.cinow.info/help?lang=en&category=faqs&item=3).

**Definitions**

* Census tracts – small, relatively permanent statistical subdivision of a county or statistically equivalent entity. Census tracts generally have a population size between 1,200 and 8,000 people, with an optimum size of 4,000 people.[[1]](#footnote-1)
* Small statistical areas – Census tracts statistically grouped together using demographic indicators.
* Dimension reduction - A method for representing a given dataset using a lower number of features (i.e. dimensions) while still capturing the original data’s meaningful properties.[[2]](#footnote-2)
* Principal component analysis – A method that reduces the number of dimensions in large datasets to principal components that retain most of the original information. It does this by transforming potentially correlated variables into a smaller set of variables, called principal components.[[3]](#footnote-3)

**Introduction**

The San Antonio metro area is consistently identified as one of the most highly economically-segregated in the country, but racial/ethnic segregation is much less pronounced. Many disparities and inequities remain hidden unless data can be broken down by both race/ethnicity and place. Data disaggregated at the census tract level or lower falls apart due to wide margins of error in survey estimates, data suppression to protect privacy, and volatile rates. ZIP codes, which are quite a bit larger than census tracts, are commonly used but are meaningless in terms of anything except efficient mail delivery. To help solve that problem, we created SSAs: clusters of census tracts grouped together based on demographic indicators. The steps and methodology below can be used to create SSAs for any geographic area that has corresponding data available at the census tract level.

**Methodology**

The basic steps of this methodology are as follows

* produce estimates at census tract level based on a specific demographic profile
* conduct principal component analysis (PCA) using demographic estimates
* spatial clustering and regionalization using PCA results

*Demographic data*

The SSAs currently used in Bexar Data Dive use the demographic indicators found in the general profile seen in Table 1 below to put together estimates that will be used for the PCA (Spielman & Folch, 2015). To collect these estimates, CINow used 2019 American Community Survey 5-year estimates at the census tract level for Bexar County, TX. CINow used the R package, Tidycensus, to pull these data using an API. The R code for pulling and calculating these estimates can be found in lines 7-229 of the “ACS\_Age\_R Code.Rmd” in our GitHub repository, named “Bexar Data Dive ACS R Code”.

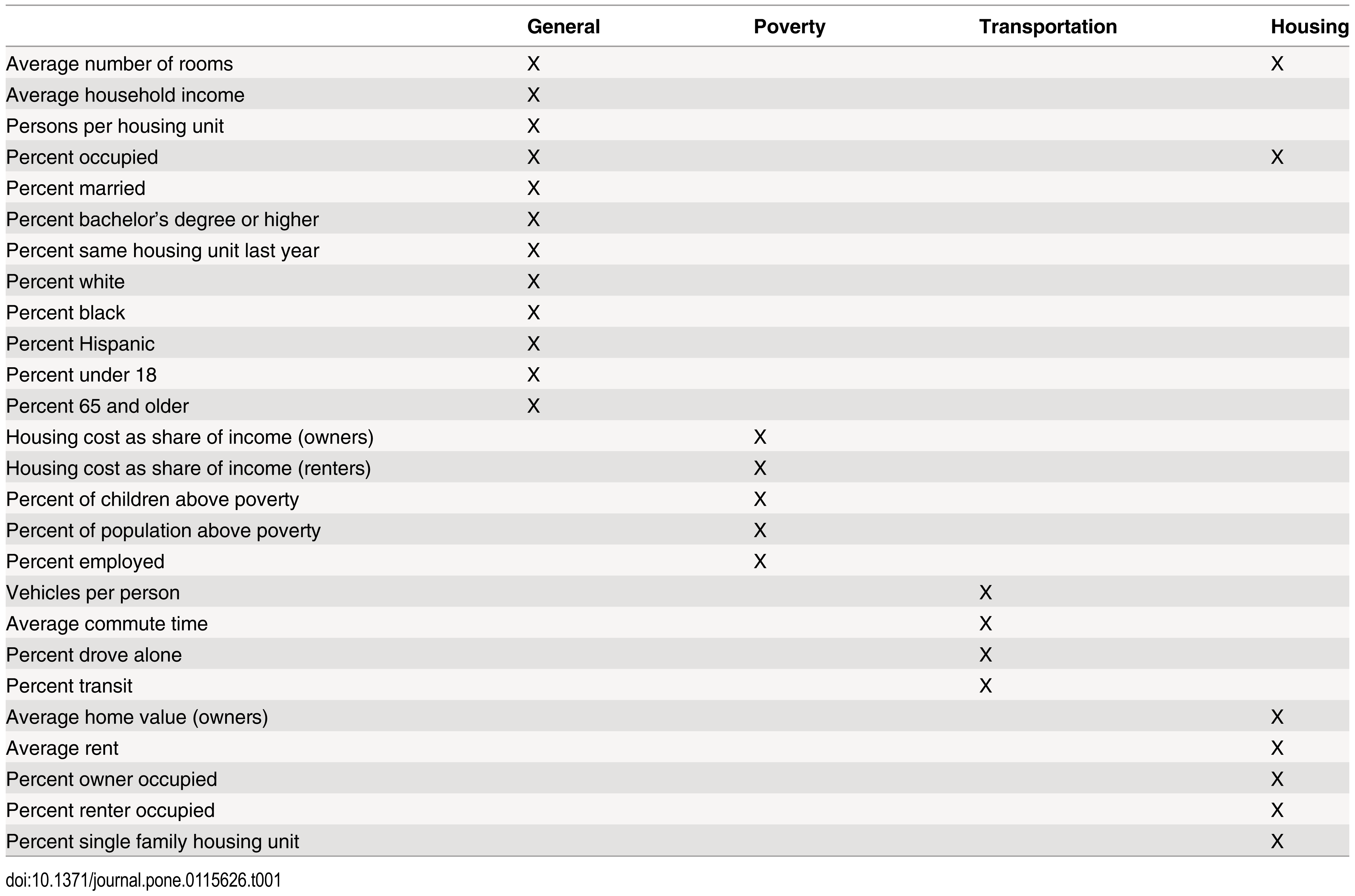


Table 1. Speilman & Folch, 2015, https://doi.org/10.1371/journal.pone.0115626

*Principal Component Analysis*

Once demographic estimates have been calculated for the census tracts. The next step is to conduct a principal component analysis (PCA) at the census tract level for dimension reduction. The goal of dimension reduction in this case is to reduce the complexity of our demographic indicators by reducing the number of features (individual indicators) while keeping the most important properties of the original data. This is done using the ‘prcomp’ function in the R ‘stats’ package. All variables need to be zeroed and standardized before running the analysis, which are options available in the ‘prcomp’ function. This prevents variables with large values, like household income, from having a larger impact on the PCA than smaller values, like those on a 0-100 scale. The R code for conducting the PCA can be found in lines 231-278 in “ACS\_Age\_R Code.Rmd” in our GitHub repository, named “Bexar Data Dive ACS R Code.”

*Spatial Clustering and Regionalization*

The final step in this process is to run a spatially constrained clustering analysis based on the principal components that explain 90% of the variance in our indicators. The SKATER (Spatial ‘K’luster Analysis by Tree Edge Removal) algorithm (Assunção et. al 2006), part of the *spdep* package in R, uses the concept of minimum spanning trees to connect all observations, weight their attribute similarities, and remove edges (boundaries) that connect observations that are not like one another. We use the SKATER algorithm because it can spatially cluster like census tracts together based on weighted attributes, our principal components, and has the option to use a queen or rook-based contiguity (point vs shared edge). Using the SKATER algorithm in the ‘sdep’ R package, the number of regions is specified by the user as well as a critical value that indicates the minimum number of units (census tracts) that should be clustered. Our final analysis created XX regions (clusters of census tracts), using a critical number of X and a rook-based contiguity. By using PCA and the SKATER algorithm, we will be able to create new regions of census tracts based on their similarities. The R code for running the SKATER function can be found in lines 281-420 in “ACS\_Age\_R Code.Rmd” in our GitHub repository, named “Bexar Data Dive ACS R Code.”

*Validation*

In order to validate our results, we chose to look at the change in statistical reliability for the Black population in census tracts versus our newly created SSAs. We used the Black population because it is a small population (7% of Bexar County) with larger margins of error at small regions like census tracts. If you choose to replicate this analysis, you can choose an indicator suited to your region to evaluate.

The measure of reliability we evaluate is the coefficient of variation (CV). The CV measures the dispersion of data points around the mean and is calculated by dividing the mean by the standard deviation. It is usually expressed as a percentage. There is no set or accepted CV value, but a lower value means a more reliable estimate. In this validation, we aim to get the CV for the Black population below 50%.

The process is outlined step by step in “ACS\_Age\_R Code.Rmd” in our GitHub repository, named “Bexar Data Dive ACS R Code.”

*Community Feedback and Bexar County Municipalities*

Based on community feedback, small municipalities in Bexar County (e.g., Alamo Heights, Castle Hills, etc.) that have their own dedicated Census Tracts were kept intact in a single SSA, rather than its boundaries straddling two or more SSAs or being lumped in with a much larger surround SSA area. The tracts for those municipalities were removed from the surrounding SSAs using GIS mapping software. Using this software municipality boundaries were overlayed with Census tracts and manually inspected one by one to identify the municipalities that could be sectioned off into their own SSA’s. When this process was complete those tracts were assigned an SSA number to represent the municipality and aggregated using the ArcGIS merge tool to create their own SSA. Some municipalities had Census tracts that did not fit within their area boundaries and therefore this process could not be performed to keep them intact in a single SSA. The final boundaries divide the county into 92 SSAs. (In comparison, the county has 76 ZIP codes.

Prepared by Cristina Martínez

1. https://www.census.gov/programs-surveys/geography/about/glossary.html#par\_textimage\_13 [↑](#footnote-ref-1)
2. Lih-Yuan Deng, Max Garzon, and Nirman Kumar, Dimensionality Reduction in Data Science, Springer, 2022. [↑](#footnote-ref-2)
3. https://www.ibm.com/topics/principal-component-analysis [↑](#footnote-ref-3)