Methodology for Election Dataset Development

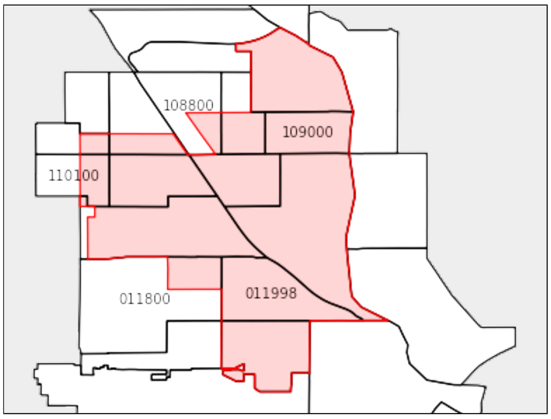
Ranked Choice Voting Project

1 Shapefiles

For each geographic feature that was not a city, county, or state, e.g. a city district, shapefiles were gathered with the locations and geometries of the feature as well as the most recent census tracts of the state containing the feature. The shapefiles for the tracts were downloaded from the Census Bureau's TIGER/Line Shapefile Web Interface, and the shapefiles for the geographic features were obtained through government websites, online public data portals, or by emailing the relevant office of the feature's administration. Some geographic features had shapefiles with coordinate systems different from that of the Census Bureau's shapefiles, and these geographies were re-projected.

2 Overlap

For each of these geographic features, it was determined which census tracts overlapped with the feature, and the percentage of each census tract included in the feature's area was calculated. The following example depicts Minneapolis's 12th city council ward outlined in red with its overlapping census tracts outlined in black, from its election in 2021.



The overlap percentages for the labeled census tracts are tabulated below as examples.

Tract Code	Overlap (%)
109000	100.00
011998	100.00
110100	44.29
108800	17.39
011800	14.88

3 ACS Data

For each election, the Census Bureau's API was used to pull data from the ACS's DP02, DP03, and DP05 tables for each tract overlapping with the election's geographic feature. For elections spanning a city, county, or state, data was pulled directly using the feature's FIPS code. Data was pulled from the 5-year estimate that best centered the election year within the 5-year timeframe. For example, 2010 elections used data from the 2012 5-year estimate.

4 Data Cleaning

Columns that consisted entirely of invalid data, which took the form of negative numbers or as the strings '(X)', '-', or 'N', were removed. Data expressed as rates or percentages were sometimes recorded across multiple variable types, so these columns were merged. Finally, redundant non-aggregatable columns, such as the percent estimate columns for variables with available estimate columns, were removed.

5 Aggregation

Because the boundaries of census tracts did not always align with the boundaries of the geographic features containing them, data was aggregated with the simplifying assumption that measured populations and characteristics are distributed evenly across a feature's area.

5.1 Aggregating Complete Tract Data

The vast majority of data were expressed as raw estimates. When this data was available for all census tracts in a geographic feature, the data were aggregated as a sum weighted by percentage overlap. The below table uses the population variable (DP05_0001E) as an example using the same election and census tracts as the map graphic on page 1.

Tract Code	**Population**	**Overlap (%)**	**Estimated Population in Feature**
109000	1812	100.00	1812.00
...	...	...	...
108800	3502	17.39	609.00
			34174.25

5.2 Extrapolating Incomplete Tract Data

In many cases, a variable was not available for some census tracts in a geographic feature. Before aggregating, the missing data was extrapolated using the tracts for which data was available. First, the frequency of the variable across the tracts for which data was available was calculated. The average of these frequencies, weighted by the estimated population of those tracts within the feature, was extrapolated to be the frequency for the unavailable tracts. An example is shown below where the number of households with a broadband Internet subscription (DP02_0154E) for the 2018 election in district 1 of San Leandro was available for 9 of the 10 census tracts in the district, unavailable only for tract 432600.

Tract	Var. Data	Population	Overlap (%)	Var. Frequency	Est. Pop. in Feature
433104	1477	4044	86.98	0.365	3517.32
...	...	...	...	...	...
432800	1256	3907	45.76	0.321	1787.65
432600	N/A	7043	78.85	N/A	5553.39

The frequency of households with broadband Internet was extrapolated as a weighted average of the frequencies for the nine available tracts and then used to estimate the number of households with broadband Internet in tract 432600:

$$\text{extrapolated frequency} = \frac{\sum(\text{variable frequency})(\text{population in feature})}{\sum \text{population in feature}} = 0.335$$

$$\text{estimated value} = (\text{extrapolated frequency})(\text{population}) = (0.335)(7043) = 2361.41$$

5.3 Aggregating Statistics

A small number of variables were expressed as statistics: percentages, means, medians, or ratios, requiring different methods for aggregation. First, a population variable was identified for each statistic variable. For means, medians, and percentages, this was the group that the statistic was calculated across. For example, the population for the unemployment rate (DP03_0009PE) was the labor force (DP03_0008E). For ratios, which usually expressed the number of males per 100 females in a group, the population was the females within that group. The table below shows examples for different types of statistics.

Label	Type	Population Label
Mean nonfamily income	Mean	Nonfamily households
Median age	Median	Total population
Unemployment rate	Percentage	Civilian labor force
Sex ratio, 65 years and older	Ratio	Females, 65 years and older

The mean was aggregated as a weighted arithmetic average of the available tracts, means with the estimated tract populations within the feature as the weights. The median was aggregated as a weighted geometric mean of the available tract medians, once again with the estimated tract populations within the feature as the weights. Like means, percentages and ratios were aggregated as weighted arithmetic averages. However, they were divided by 100 before aggregation and multiplied by 100 after aggregation to account for their representation as 100 times a proportion.