New York University - Center for Urban Science and Progress

California Dream Index

Visualizing and Analyzing Economic Mobility in California

Michael Carper and Aren Kabarajian
Dr. Eric Corbett, Postdoctoral Associate
Patrick Atwater, Senior Research Analyst at California Forward

Abstract

The California Dream Index (CDI) was started by the nonprofit organization California Forward[1] to provide policymakers and other stakeholders with an understanding of socioeconomic mobility around the state. The CDI currently stands as a web tool displaying regional and county-level statistics on ten indicators of economic mobility. At county granularity, the tool lacks the precision necessary to track the economic livelihood of specific neighborhoods and smaller areas. The capstone team addressed this issue by re-creating the CDI's interactive state map visualizations at census tract granularity. The completion of this goal will provide local level stakeholders with the power of tangibility when lobbying to improve areas of economic vulnerability. In a second phase of the project, the capstone team performed a series of California resident clustering analyses using some of the same indicators scored in the map visualizations as features. The first of these, a comprehensive analysis of residents in 2019, revealed groups of economic archetypes based on eight of the CDI indicators. The second clustering exercise was a series of analyses performed separately on each decade from 1990 to 2019, relying only on housing, education and commute times as features. The team hypothesized that cluster scores on the selected indicators of economic mobility would all correlate positively with cluster incomes. While some of the selected features proved to be strong, consistent indicators of economic status, others displayed more nuanced relationships with income than expected.

[1] Home - California Forward, Nonprofit Leading in Economy and Policy (cafwd.org)

1. Introduction

Economic inequality is a worsening issue in the United States, and despite tax policies and safety net programs, the state of California is no exception. Part of this issue of inequity is demonstrated in the metric of income. According to the Public Policy Institute of California, the divide between the state's highest earners and the rest of the population is increasingly concerning each decade.[2] The 90th percentile of household income has increased by 60% since 1980, while median and 10th percentile household incomes have increased by merely 24% and 20%, respectively. Wealth inequality is also a significant issue, as PPIC reports that 20% of the state population's net worth is held by just 2% of California residents.

Addressing inequality or any issue with an economy should begin with a clear, detailed understanding of its present and historical status. Metrics like gross domestic product (GDP) and unemployment figures are frequently used on their own to describe the health of economies. In the US, individual state economies are often evaluated by whether or not their budget is balanced or in deficit, known by some as budgetism[3]. When evaluating economic activity and the livelihoods of citizens, employing any of these metrics by themselves results in a drastically oversimplified snapshot that neglects a depiction of the majority of the population's financial satisfaction, security or the level of opportunity for those to change between generations.

The purpose of the California Forward-sponsored California Dream Index[4] (CDI) is to provide a comprehensive measurement of the health of the state's economy and the livelihood of its citizens by scoring ten indicators of economic mobility as well as a composite index called the Overall CDI Score. The CDI and this capstone team's contributions are intended to supplement oversimplified metrics of the economy and provide residents and policymakers with a comprehensive understanding of inequality, economic mobility and general livelihood in areas across the state. As it stands, the CDI includes a web tool with an interactive geographic visualization of regional and county-level scores for the ten indicators and overall index. These are available on the CDI website for years 2010 to 2019 and are developed mainly using open data from the American Community Survey.

[2] Bohn & Thorman, "Income Inequality in California." *Public Policy Institute of California*. Jan. 2020.
[3] Matthews, "California: Success Isn't Measured by Budget Deficit or Surplus." *Zocalo Public Square*. 24 Jan., 2021.
[4] CA Dream Index - California Forward - Tracking Our Progress (cafwd.org)

California Forward used a study from Raj Chetty[5] on the geography of intergenerational mobility as a guide to select ten CDI indicators to serve as critical descriptors of an area's economic status and the livelihood and mobility of its residents. The literature pointed to income equality, primary and post-secondary education, social capital, family stability and a lack of residential segregation of race and ethnicity as key determinants of economic mobility. The indicators include metrics on these factors and more. See Table 1 for the full list.

Pinpointing specific areas and populations within California that need economic assistance is not possible with the current county-level granularity on the California Dream Index web tool. California is home to almost 40 million residents across more than 160,000 square miles, and there are only 58 county boundaries. This inhibits the ability to dissect economic prosperity in cities and towns and petition for change to policymakers. The first goal of this capstone project is to re-create the CDI visualizations at the census tract level, creating profiles of economic mobility for neighborhoods and areas within US Census-defined tract geographies across the state. These more granular interactive maps will allow California Forward to inform decision makers and policymakers on a more local level, while also painting a finer picture of the state's economic status at large.

In addition to the CDI's current lack of localized geographic granularity, the web tool also lacks an analytical component to complement its geographic features. Thus, the second goal of this capstone project is to provide a conception of resident archetypes based on some of the same economic indicators used in the map scoring. To accomplish this, the team first performed a clustering analysis on 2019 ACS microdata accessed through IPUMS USA[6], creating economic archetypes of current residents based on eight of the CDI indicators as features. The team hypothesized that cluster performance in these eight features would correlate positively with cluster median incomes. This analysis is intended to reveal subpopulations who have similar attributes in terms of present economic livelihood and mobility for the future.

In a second analysis exercise, clusters have been developed and compared between each decade from 1990 to 2019, again using the IPUMS microdata archives. Due to availability in IPUMS datasets before 2010, only three features were used for this analysis: housing, education and commute time. This analysis is intended to provide clarity on any temporal progression and

[5] Raj Chetty et al., "Where is the Land of Opportunity?" *NBER Working Paper No. 19843*, Jan. 2014.
[6] IPUMS USA

regression of Californians' economic mobility and also provide insights on how the indicators' connections with income have evolved over time. Both clustering analyses also demonstrate the connection between the CDI indicators with economic status. The team hypothesized that the cluster makeups and feature scores would differ noticeably from each decade to the next because of the constantly evolving economy in California and the United States as a whole.

2. Data

2.1 Visualization Exercise Data

The ten CDI indicators of economic mobility and their sources are listed in Table 1.

Indicator	***Sourcing***
College and CTE Credentials	US Census American Community Survey
Early Childhood Education	US Census American Community Survey
Income above cost of living	US Census ACS & United Way Real Cost Measure
Rent Burden	US Census American Community Survey
Home Ownership	US Census American Community Survey
Broadband Access	US Census American Community Survey
Short Commutes	US Census American Community Survey
Prosperous Neighborhoods	US Census American Community Survey
Clean Drinking Water	California State Water Resources Control Board database
Air Quality	CalEnviroScreen

Table 1. CDI Indicator Sources

Most of the data used for the metrics of economic mobility are sourced from the United States Census Bureau, with a few exceptions for niche indicator categories. Before the capstone team began this project, state, region, county and tract-level scores for each of these attributes had been pre-compiled by the California Forward staff in tables with geography codes. The files are stored in .csv files on Dropbox. Each indicator score is stored here in its raw calculated form. For example, the Air Quality scores here are average AQI readings over the course of a year. On the interactive map graphs, raw scores are presented. All scores are normalized onto a [0, 100] interval for aggregation into the overall score and for the site's multi-line graphs[7] that provide

[7] Geography Detail | California Dream Index (cadreamindex.org)

temporal views of all indicators at once. Overall CDI scores are calculated by averaging the ten indicators' normalized scores and multiplying by 100. Normalization is performed using the following formula, where Min and Max refer to the statewide minimum and maximum raw scores for that indicator:

$$Score \ = \ [Raw\,Value \ - \ Min] \ * \ 100 \, / \, (Max \ - \ Min)$$

Census tract granularity was chosen for the visualization component because of its compatibility with pre-existing California Forward-sourced data and the fact that tracts align with county geographies, which are currently viewable on the CDI web tool.

2.2 Clustering Analysis Data

2.2.1 Comprehensive 2019 Clustering Data

For the 2019 Comprehensive clustering analysis, the team utilized census microdata from IPUMS USA. This data contains similar information to the geographic data used for the visualizations but on a resident and household level as opposed to a tract or county level. Desired variables and years are selected, then datasets are requested for download from IPUMS on their website. The first clustering exercise employs 2019 5-Year ACS estimate data, which has a 5% sampling density. While only four of the CDI indicators could be calculated with this data, 40 variables were selected for this dataset to allow for rich post-analysis insights on cluster demographics. The team merged this dataset with four county-level features from the previously described Dropbox files. These four were selected because of their applicability to all or most of the residents within these geographies. County was the selected geographic distinction because county IDs are available in the IPUMS data while census tract IDs are not. All features employed for this clustering analysis, their post-processing value descriptions and their sources are listed in Table 2.

Feature	Description	Source
Housing (categorical)	0 = Renter paying >=30% of their income on rent 1 = Renter paying <30% 2 = Owner	2019 ACS 5yr microdata Calculated by the team using the following IPUMS variables: Household income Home ownership (binary)

		Gross monthly rent
Education (binary)	0 = Less than 2-year college degree or equivalent 1 = 2-year college degree or higher education level	2019 ACS 5yr microdata
Broadband Availability	10 = Broadband-serviced household 20 = Non-broadband-serviced household	2019 ACS 5yr microdata
Commute Time	Number of minutes spent in daily commute to work	2019 ACS 5yr microdata
Drinking Water Quality	% of residents not living near an area with drinking water complaints	CDI Dropbox files (county level)
Air Quality	Average EPA Air Quality Index	CDI Dropbox files
Early Education	% of 3-4 year olds enrolled in preschool	CDI Dropbox files
Prosperous Neighborhoods	% of county residents who do not live in census tracts of concentrated poverty	CDI Dropbox files

Table 2. 2019 Comprehensive Clustering Feature Descriptions

2.2.2 Multi-Year Clustering Data

For the Multi-Year clustering exercise, county-level values were not included because California Forward's Dropbox data files only date back to 2010. The features listed in Table 3 were employed for separate analyses on 1990, 2000, 2010 and 2019 census microdata with consistent sampling density.

Indicator	Description	Source
Housing (categorical)	0 = Renter paying >=30% of their income on rent 1 = Renter paying <30% 2 = Owner	1990 5% state 2000 5% 2010 ACS 5yr - 5% density 2019 ACS 5yr - 5% density
Education (binary)	0 = Less than 2-year college degree or equivalent 1 = 2-year college degree or higher education level	1990 5% state 2000 5% 2010 ACS 5yr - 5% density 2019 ACS 5yr - 5% density

Commute Time	Number of minutes spent in daily commute to work	1990 5% state 2000 5% 2010 ACS 5yr - 5% density 2019 ACS 5yr - 5% density

Table 3. Multi-Year Clustering Feature Descriptions

3. Methodology

3.1 Visualization Methods

The first goal of the visualization phase was to prototype tract-level interactive map visualizations on the Lake Tahoe area and present findings to Tahoe Prosperity Center[8]. Interactive map graphs were created for each CDI indicator and the composite scores, and for this Tahoe trial only, line graphs were created to depict the evolution of normalized indicator scores from 2010 to 2019.

All geographic plots were constructed using Datawrapper, which was chosen due to its ease of sharing, editing and its element of interactivity. The team initially built individual line graphs for each indicator in each tract using Matplotlib. Seeking a more appealing format, the team then visualized all indicators together on interactive multi-line graphs for each tract using Datawrapper[9]. Data was prepared and processed for these exercises manually within Excel. After the Tahoe exercise was completed, the team moved on to create statewide visualizations.

The Tahoe area includes 17 US Census tracts, while the state of California has over 8000, so the team shifted data preparation and processing methods to Python to automate the process for the increased scale. The team wrote a function that streamlined the process of trimming and cleaning the statewide Dropbox files provided by California Forward into usable formats, extracting the geographic FIPS Code, census tract name, year and indicator score desired. Another function was written and utilized to create “difference” dataframes and maps, subtracting all tract scores of one year from another.[10] Interactive statewide maps were created again with Datawrapper, covering all indicator scores for 2010, 2019 and the decadal differences between the two[11] in each tract.

[8] Tahoe Prosperity Center l Uniting Tahoe's Communities

[9] See Figure 2 in Appendix 7.3

[10] See Appendix 7.1 for team GitHub repository

[11] See Figures 4, 5 in Appendix 7.3

3.2 Clustering Analysis Methods

The team first requested and downloaded a single dataset from IPUMS that included all the necessary variables and samples to perform both sets of clustering analyses. The cleaning, processing, clustering and interpretation stages were all performed similarly among the 2019 Comprehensive and Multi-Year analyses. One step taken in the data cleaning and processing stage was creating a custom housing feature using three IPUMS variables. Rows were first separated by homeowners vs. non-owners, then non-owners' gross rent cost was divided by their household income to determine whether they were paying less than 30% of their income on rent, creating a custom ranked housing variable. This was done to remain as consistent as possible with the original CDI indicator calculations. See Table 2 for further details.

The team decided to use a K-means clustering approach for all analyses because the method weights each feature equally as long as the data is standardized. A silhouette score loop was employed to identify the ideal number of clusters for each analysis. Strong scores were achieved at seven clusters for the 2019 Comprehensive and four clusters for each decade of the Multi-Year. While some scores for higher cluster counts slightly exceeded the chosen count scores, the team believed that these choices were optimal for result interpretation.

In order to effectively use variables with values on different scales, the data was standardized with a standard scaler before the clustering step and then reverse-transformed afterwards to retrieve raw values for interpretation of the results. For the 2019 Comprehensive analysis, eight features were used to create the seven clusters. For the Multi-Year, datasets were separated by year (1990, 2000, 2010, 2019), and the same three features were used to create separate sets of four clusters for each year to be compared over time. See Tables 2 and 3 for details on these features.

After the analysis, the household weight census variable was used as a multiplier for each row, shifting the dataset from household level to resident level as intended. Results are depicted as tables and histograms revealing cluster feature distributions across all indicators, race and ethnicity makeups, county locations of cluster residents (Multi-Year only), household income and resident ages. For a simplified look at the cluster makeups, ranking tables were produced and developed into radar charts to demonstrate connections between performance in the features and income.

3.3 Study Limitations

The team faced a predicament in the data cleaning stage of the clustering analyses. For the commute time variable, roughly 500,000 rows had 0 values for daily minutes spent commuting to work. It can be assumed that some of these are due to unemployment, but the rate is roughly 8% in California, which would not account for nearly one third of the entire dataset. Another explanation could be those who work from home, but that amount seems unlikely. Regardless, the team did not want to lose all of these rows, so they were kept and included in the analysis as they were.

Another limitation with the census microdata was the lack of the broadband variable 2015. Coupled with the fact that the CDI geographic data only goes back to 2010, the team was forced to limit the Multi-Year analysis features to just three (see Table 3). Also, the mixture of ordinal, binary and numerical features in the data presented difficulty in both modeling and interpretation. The team found that the traditional visualization method of scatter plots was uninterpretable for these analyses due to the mixed variable types as well as the sheer number of rows in the data.

A limitation to the actual analysis was the selection of the clustering method, as the distance driven method by K-means may not have been a perfect fit for the 2019 Comprehensive clustering. The team attempted and rejected the use of hierarchical clustering but did not have time to test the Gaussian Mixture method.

A general limitation to the use of census data is the potential bias that a census holds and the fact that it might not depict a true composition of the population. Recently the census count has been fairly accurate at the national level, with the total 2010 undercount at 0.01%, but equal representation among subpopulations is a concern[12]. This is revealed in the post-enumeration survey of the census which investigates missing people in the count. For example, in 2010 the Non-Hispanic Black population had 2.06% missing, while the Non-Hispanic White population had only a 0.83% missing count. This disparity could have implications towards policy decisions in the long term. For this project's purposes, census data and microdata was the most detailed and

[12] "Census Accuracy and THE Undercount: Why It Matters; How It's Measured." *Census Counts*, 2 Jan. 2020, https://censuscounts.org/whats-at-stake/census-accuracy-and-the-undercount-why-it-matters-how-its-measured/

comprehensive available, but using them was a decision made with these limitations in mind. For additional limitations of this study, refer to Appendix section 7.2.

4. Results

4.1 Visualization Exercise Results

The final result of the visualization project component was a set of twenty four interactive state maps that showed each tract's indicator scores for 2019, 2010 and the difference between the two years. These Datawrapper plots highlighted differences in mobility among areas across the state while also revealing changes over the decade. The most prominent indicator changes from 2010 to 2019 were evident in the Early Childhood Education map and Affordable Rent map[13]. The overall visual trend for childhood education across the state was an increase over the decade. There are also numerous areas where the early education rate decreased, and overall, significant changes in either direction were apparent across the state. Increased funding in California Head Start, a program to help low-income children with resources, could be attributed to some of the positive change happening in minority dominant neighborhoods.[14]

Changes in Affordable Rent over the decade were not as stark, and the average across all tracts showed a positive increase in renters paying less than 30% of their income. This aligns with historical data for California, as median rent as a share of income decreased over the decade[15]. Looking at gross rent cost alone could be misleading, as the state median increased by $300 from 2010 to 2019. Insights like these and a closer look at neighborhoods and towns around the state are now visible with the capstone team's maps. These will boost California Forward's internal understanding of how economic health has changed over this decade and will be an informational asset when working with other organizations. Links to all these maps are available on a document that is linked in Appendix section 7.1.

4.2 Clustering Analysis Results

[13] See Figures 6, 7 in Appendix 7.3

[14] Collier, Michael. "Head Start Programs in California Rebound as Funding Increases." EdSource, edsource.org/2015/head-start-programs-in-california-rebound-as-funding-increases/74901

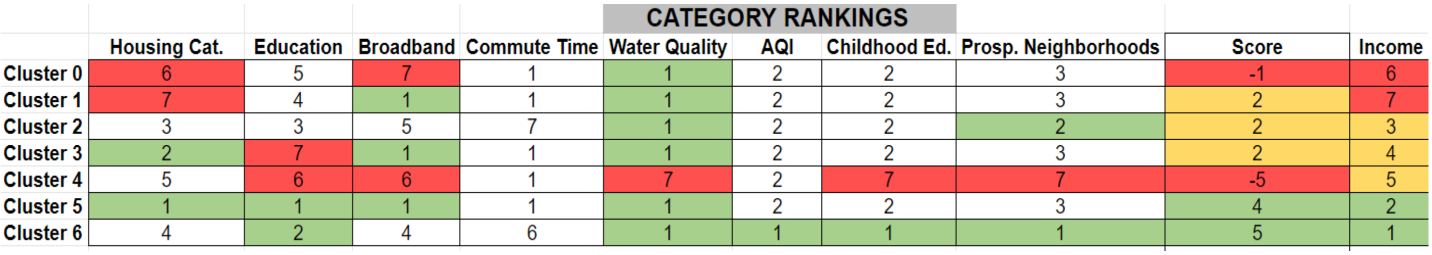
[15]"Residential Rent Statistics for California | Department of Numbers." Www.deptofnumbers.com, www.deptofnumbers.com/rent/california/. Accessed 8 Aug. 2021.

The initial hypothesis of both clustering analyses was that there would be a direct positive correlation between cluster income and performance on each indicator. For example, if a cluster has highly rated features like housing, education and broadband access, the median income of residents making up that cluster would be higher than clusters with low-ranking features. This prediction was made based on the confidence of the capstone team and sponsor organization that the CDI indicators are predictors of economic welfare.

4.2.1 Comprehensive 2019 Clustering Results

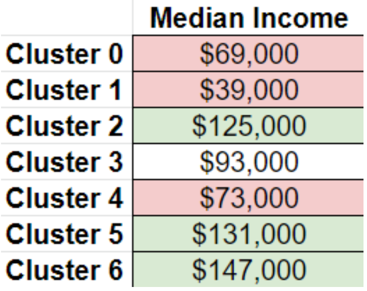
The 2019 clustering presented very detailed findings due to its eight features and seven different clustering groups. The choice of seven as the cluster count led to distinguishable differences between each of the groups. Nearly every cluster has at least a $15,000 difference in median household income from all others. In order to simplify interpretation, a scoring method was used for each cluster that added one to its score if it was in the top two ranks of a feature and subtracted one if it was in the bottom two. Features with similar scores across the board were partially omitted from the scoring.

The clusters with the two highest incomes in the analysis indeed had the top two overall scores. Both were also the highest ranking in terms of college attainment and early childhood education, and the highest earning cluster had a significantly better air quality average compared to others. These highest income clusters also had high home ownership rates (over 65%) and together combined for close to 40% of the sample size.



					CATEGORY RANKINGS					
	Housing Cat.	Education	Broadband	Commute Time	Water Quality	AQI	Childhood Ed.	Prosp. Neighborhoods	Score	Income
Cluster 0	6	5	7	1	1	2	2	3	-1	6
Cluster 1	7	4	1	1	1	2	2	3	2	7
Cluster 2	3	3	5	7	1	2	2	2	2	3
Cluster 3	2	7	1	1	1	2	2	3	2	4
Cluster 4	5	6	6	1	7	2	7	7	-5	5
Cluster 5	1	1	1	1	1	2	2	3	4	2
Cluster 6	4	2	4	6	1	1	1	1	5	1

Table 4. 2019 Comprehensive Clustering Category Rankings



	Median Income
Cluster 0	$69,000
Cluster 1	$39,000
Cluster 2	$125,000
Cluster 3	$93,000
Cluster 4	$73,000
Cluster 5	$131,000
Cluster 6	$147,000

Table 5. 2019 Comprehensive Clustering Median Incomes

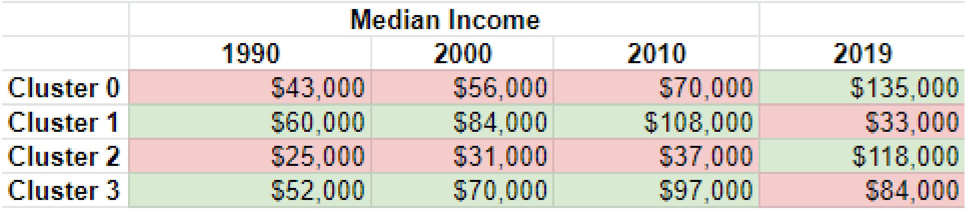
The remaining clusters, especially on the low-ranking end, had a more ambiguous overall correlation when it came to the scoring method. The cluster with the lowest household income had a mid-ranked rate of college education and was tied for the highest ranking of broadband access. This shows a potential flaw in assessing each of the indicators equally, as although broadband access is the highest here, most clusters had a high percentage of access as well. The features did not provide a perfect correlation with income, as the cluster ranking fifth of the seven groups for income had by far the lowest overall feature score. However, housing proved to hold as a strong predictor at the lower end of the income rankings, as the two lowest income clusters had the least homeowners and most renters paying over 30% of their income.

There was certainly a connection between feature performance and income, but the team's hypothesis must be rejected, as the features were not seamless predictors for lower income clusters. This points to a conclusion that lower income groups' distinct social and economic struggles are more nuanced, making it harder to address needs without a more detailed analysis of their resources. Listed in Appendix section 7.4 are qualitative cluster summaries and radar charts depicting polygons that represent each cluster's ranking makeup. These depictions of the results provide rich demographic insights.

4.2.2 Multi-Year Clustering Results

In addition to the overarching clustering hypothesis, for the Multi-Year clustering, the team predicted that the cluster makeups and feature scores would differ noticeably between each decade due to economic changes over each decade in California. An additional reason for this prediction was that each year was put into a clustering algorithm individually. Results showed that cluster makeups for years 1990, 2000 and 2010 were very similar, rejecting this prediction. For these three decades, the cluster composition remained largely the same except for a change in housing ranking from 1990 to 2000. Also, the lowest income group decreased in size from 2000 to 2010.

In 2019, a complete shift in cluster rankings occurred, suggesting an overall shift in the California economy over this decade. Housing ownership proved to be more predictive of income in 2019 than before. Education maintained a strong, direct correlation with income throughout all four decades, which aligns with the main clustering hypothesis. The commute time feature, however, showed a consistently strong negative correlation with income, which was contrary to the team's expectation. After further investigation, this negative correlation is likely attributed to the number of zero values for commute time in this dataset (see Section 4).



	Median Income			
	1990	2000	2010	2019
Cluster 0	$43,000	$56,000	$70,000	$135,000
Cluster 1	$60,000	$84,000	$108,000	$33,000
Cluster 2	$25,000	$31,000	$37,000	$118,000
Cluster 3	$52,000	$70,000	$97,000	$84,000

Table 6. 2019 Multi-Year Clustering Median Incomes

The Multi-Year clustering revealed key insights into changes in housing importance and the wealth gap. The housing category factor was more weakly correlated with income in 1990 and consistently became more positively correlated with income each decade. The highest income group had the third best housing rating in 1990, the second best in 2000 and 2010, and the highest rating in 2019. The wealth gap between the highest and lowest income groups also grew over this time period by close to twenty percent, showing wealth concentration increasingly moving to the hands of the elite. These changes over the last few decades prove that new methods of intervention will be necessary to improve economic equality in California. For further results and insights, as well as maps of the county locations of residents within each cluster, refer to Appendix section 7.5.

5. Conclusion

The clustering analyses and visualizations produced by this capstone team will help California Forward and any stakeholders they work with to analyze and contextualize economic mobility and economic archetypes in California. The visualizations created a granular, local level conception of indicator performance across the state over the last decade. These will mostly be used by the organization internally, but replicating the team's methodology with Leaflet, the CDI website's visualization platform, would provide an improved web tool for public use.

While the team's main clustering hypothesis was rejected, some indicators proved to be strong predictors of economic status. To build on this analysis, alternate clustering methods such as a Gaussian Mixture could be attempted and additional data could be supplemented for increased validity, especially with commute times. The code repositories and methodologies of this capstone project have been passed on to California Forward, allowing them to easily replicate and build upon these exercises as future years of census data are released. With these deliverables, the organization can continue to supplement oversimplified metrics and strive for economic equity and livelihood in the state of California.

6. References

[1] *California Forward*, https://cafwd.org/

[2] Bohn & Thorman, "Income Inequality in California." *Public Policy Institute of California*. Jan. 2020. https://www.ppic.org/publication/income-inequality-in-california/.

[3] Matthews, "California: Success Isn't Measured by Budget Deficit or Surplus." *Zocalo Public Square.* 24 Jan., 2021. https://www.zocalopublicsquare.org/2021/01/26/california-budget-deficit-covid-19/ideas/connecting-california/

[4] *California Dream Index*. https://cafwd.org/california-dream-index/

[5] Raj Chetty et al., "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States". *National Bureau of Economic Research*, Jan. 2014. https://www.nber.org/system/files/working_papers/w19843/w19843.pdf

[6] Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek. *IPUMS USA*: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://usa.ipums.org/usa/index.shtml

[7] **"**Geography Detail", *California Dream Index.* https://www.cadreamindex.org/geography/fresno-county/

[8] *Tahoe Prosperity Center*. https://tahoeprosperity.org/

[9] "Census Accuracy and THE Undercount: Why It Matters; How It's Measured." *Census Counts*, 2 Jan. 2020.

https://censuscounts.org/whats-at-stake/census-accuracy-and-the-undercount-why-it-matters-how-its-measured/

7. Appendix

7.1 Project Links

GitHub Repository:
https://github.com/aren-kab/CDI-Capstone
All Datawrapper Statewide Maps:
https://docs.google.com/document/d/1-OEWw4YZRT--lM-cr1LsP945NfL_4M_ver1qHMoB0CA/edit

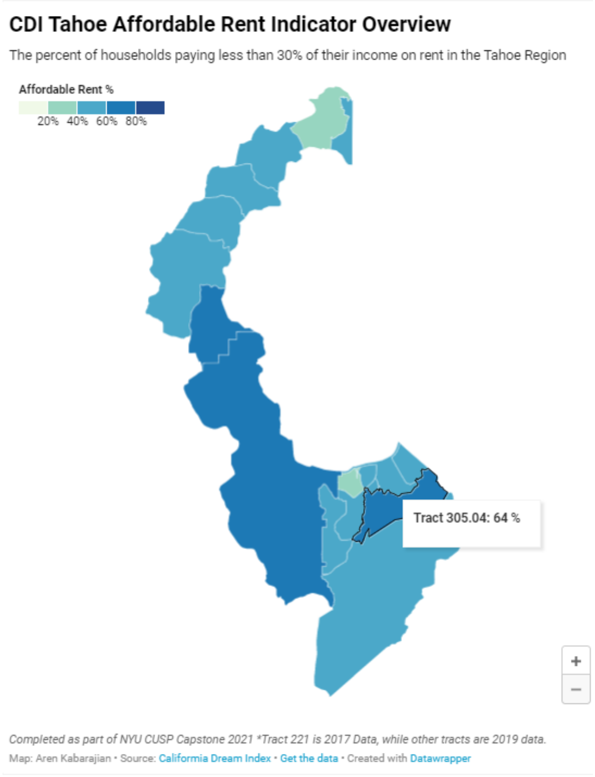
7.2 Additional Limitations

In the tract-level visualization phase, the most significant limitation was missing values in the Dropbox data files. Due to time constraints and the relatively small share of missing values in comparison to available ones, the capstone team made no attempt to fill in these gaps. Instead, geographic areas with N/A values in a certain category were depicted as zero values and a caveat for this was added to each graph's subtitle.

The clustering analysis presented some difficulty with data cleaning and processing, and some assumptions were made. Rows with age values below 21 were removed for sensibility with the education and rent cost features. This eliminated roughly 400,000 rows out of the original 1.8 million of the 2019 portion of the dataset. Rows with N/A values for the broadband availability and home ownership census variables were dropped from the whole dataset. This cost the dataset about 180,000 rows in the 2019 set but was necessary for a clean look at their "yes" and "no" binary values. The number of N/A values were similar for other years and approached in the same way. In the process of merging the 2019 Comprehensive part of this dataset with county-level values, roughly 60,000 rows were lost due to some invalid county FIP codes.

Overall, this project provides meaningful insights but is not intended to act as a standalone depiction of Californians' economic mobility. The visualizations and clustering results are products of complex data assemblages that include biases within the data and assumptions made throughout the team's procedure. The team acknowledges that this project provides a subjective glimpse at inequity and should act as a supplement to other references and metrics but not necessarily replace or overturn them.

7.3 Visualization Examples

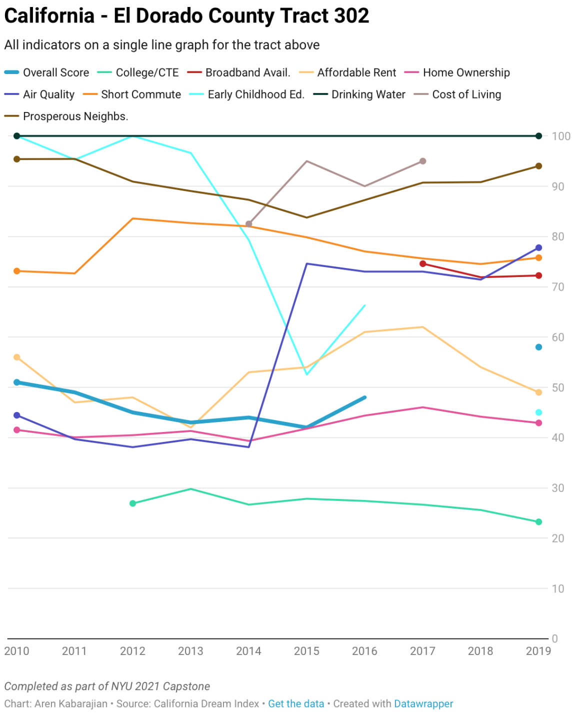

CDI Tahoe Affordable Rent Indicator Overview

The percent of households paying less than 30% of their income on rent in the Tahoe Region

Completed as part of NYU CUSP Capstone 2021 *Tract 221 is 2017 Data, while other tracts are 2019 data.

Map: Aren Kabarajian • Source: California Dream Index • Get the data • Created with Datawrapper

Figure 1. Tahoe Affordable Rent Map

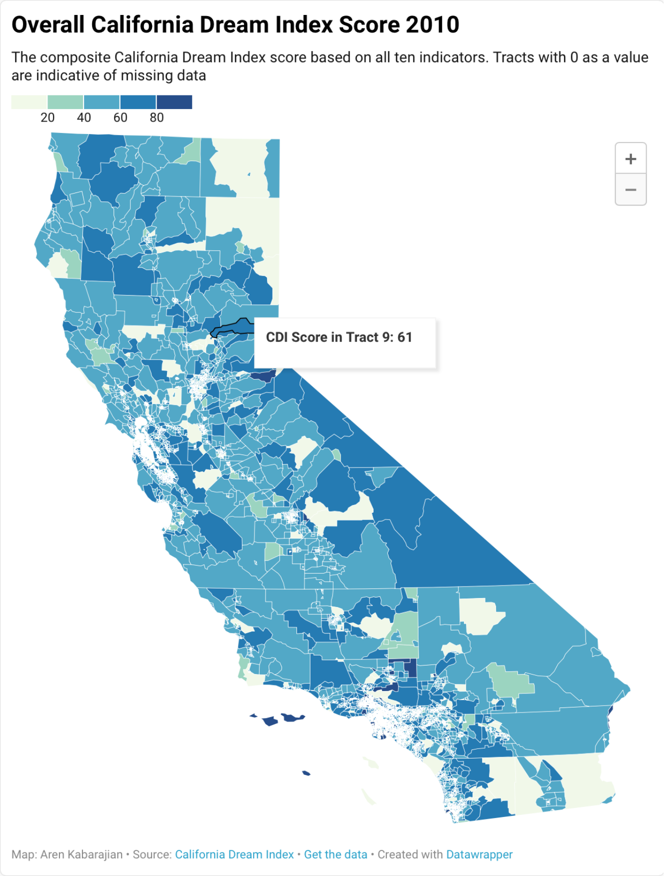

California - El Dorado County Tract 302

All indicators on a single line graph for the tract above

Completed as part of NYU 2021 Capstone

Chart: Aren Kabarajian • Source: California Dream Index • Get the data • Created with Datawrapper

Figure 2. Tahoe Multi-indicator Chart

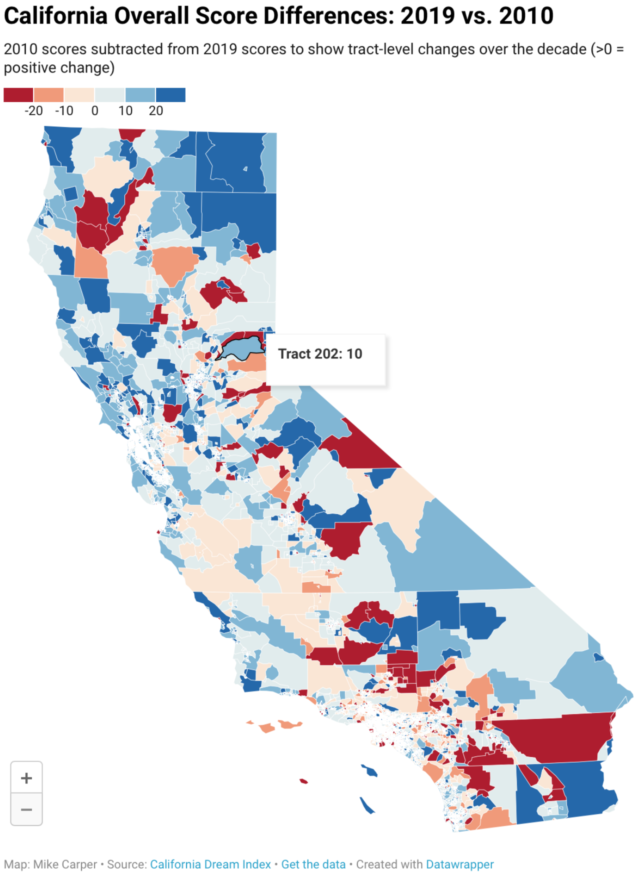


Overall California Dream Index Score 2010

The composite California Dream Index score based on all ten indicators. Tracts with 0 as a value are indicative of missing data

Map: Aren Kabarajian • Source: California Dream Index • Get the data • Created with Datawrapper

Figure 4. CDI Score Map 2010

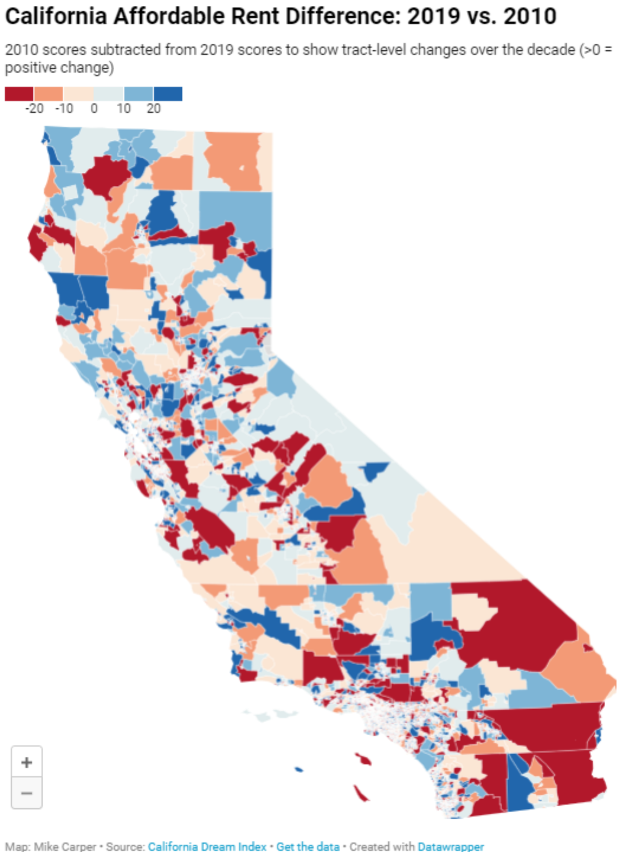


California Overall Score Differences: 2019 vs. 2010

2010 scores subtracted from 2019 scores to show tract-level changes over the decade (>0 = positive change)

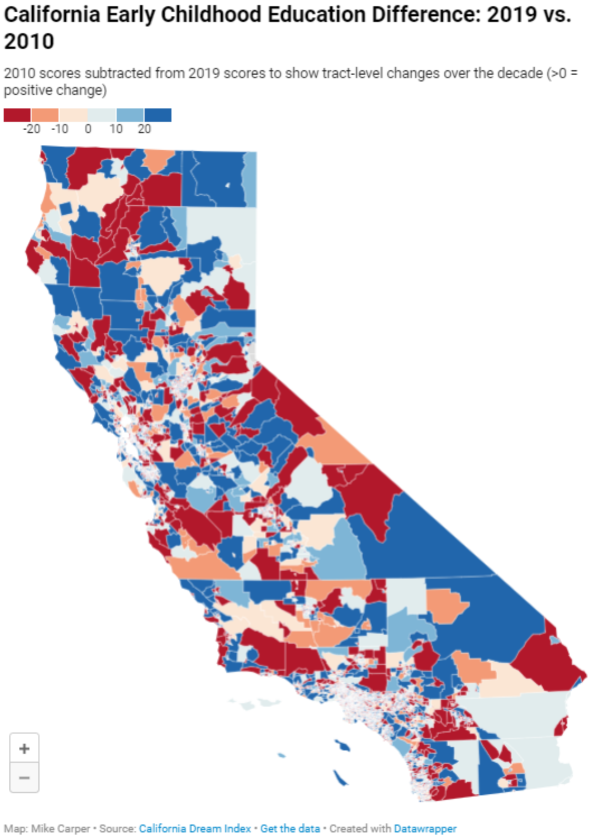
Map: Mike Carper • Source: California Dream Index • Get the data • Created with Datawrapper

Figure 5. CDI Score Difference 2019 - 2010


California Affordable Rent Difference: 2019 vs. 2010

2010 scores subtracted from 2019 scores to show tract-level changes over the decade (>0 = positive change)

Map: Mike Carper • Source: California Dream Index • Get the data • Created with Datawrapper


California Early Childhood Education Difference: 2019 vs. 2010

2010 scores subtracted from 2019 scores to show tract-level changes over the decade (>0 = positive change)

Map: Mike Carper • Source: California Dream Index • Get the data • Created with Datawrapper

Figure 6. Affordable Rent Difference

Figure 7. Early Childhood Education Difference

7.4 Comprehensive 2019 Clustering Results

2019 Comprehensive Cluster Category Rankings:

	Housing	*Education*	*Broad-band*	*Commute*	*Water Quality*	*AQI*	*Childhood Ed.*	*Neighb-orhoods*	*HH Income*
Cluster 0	6	5	7	1	1	2	2	3	6
Cluster 1	7	4	1	1	1	2	2	3	7
Cluster 2	3	3	5	7	1	2	2	2	3
Cluster 3	2	7	1	1	1	2	2	3	4
Cluster 4	5	6	6	1	7	2	7	7	5

Cluster 5	1	1	1	1	1	2	2	3	2
Cluster 6	4	2	4	6	1	1	1	1	1

Comprehensive 2019 Cluster Summaries:

Cluster 0: medium-sized group of less educated, housing-burdened, broadband-neglected Hispanic-heavy lower-middle class

Cluster 1: medium group of less educated, severely housing-burdened Black and Hispanic-heavy lower-middle class; lowest income cluster

Cluster 2: small group of hard working* racially mixed upper-middle class

Cluster 3: large group of least educated Hispanic-heavy middle class

Cluster 4: small group of less educated White-dominant, Native American and Hispanic-heavy lower middle class living in less prosperous neighborhoods with suffering water quality andless access to childhood education

Cluster 5: large group of well-educated Asian-heavy upper-middle class

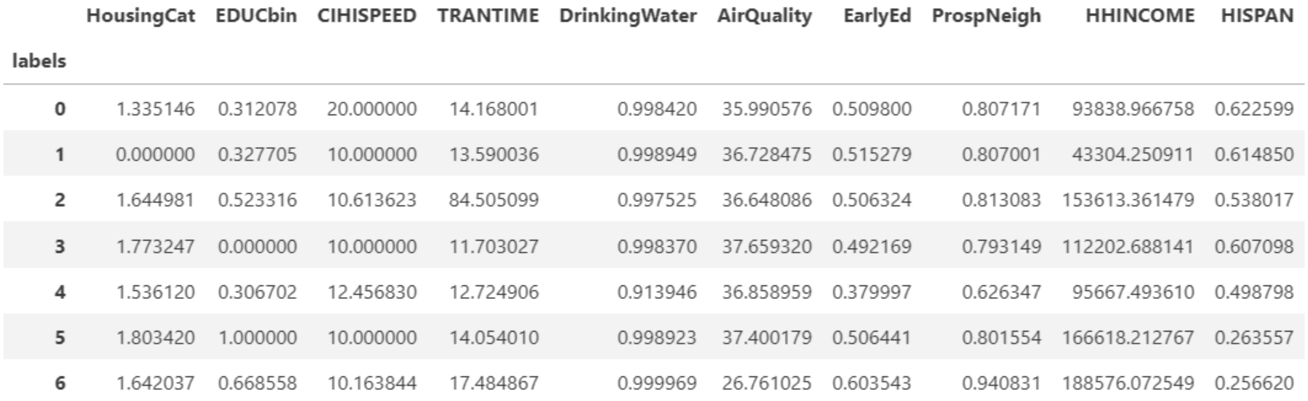
Cluster 6: medium group of well-educated Asian-heavy upper-middle class living in more prosperous neighborhoods with better air quality and access to childhood education; highest income cluster

** "hard working" = significantly longer median commute time than other clusters*

2019 Comprehensive Cluster Sizes:

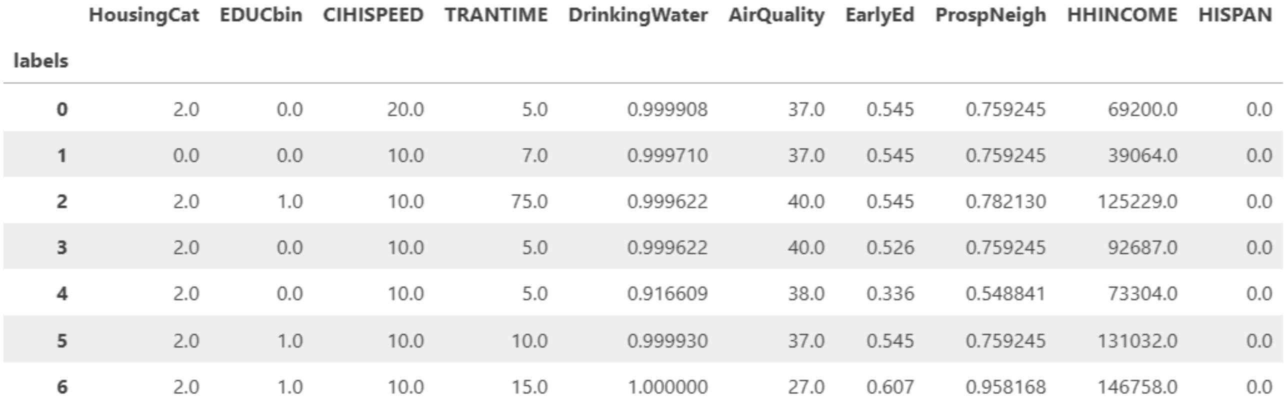
Cluster	**Size**
0	12% of total
1	12%
2	5%
3	24%
4	5%
5	21%
6	17%

2019 Comprehensive Cluster Feature and Income Means:



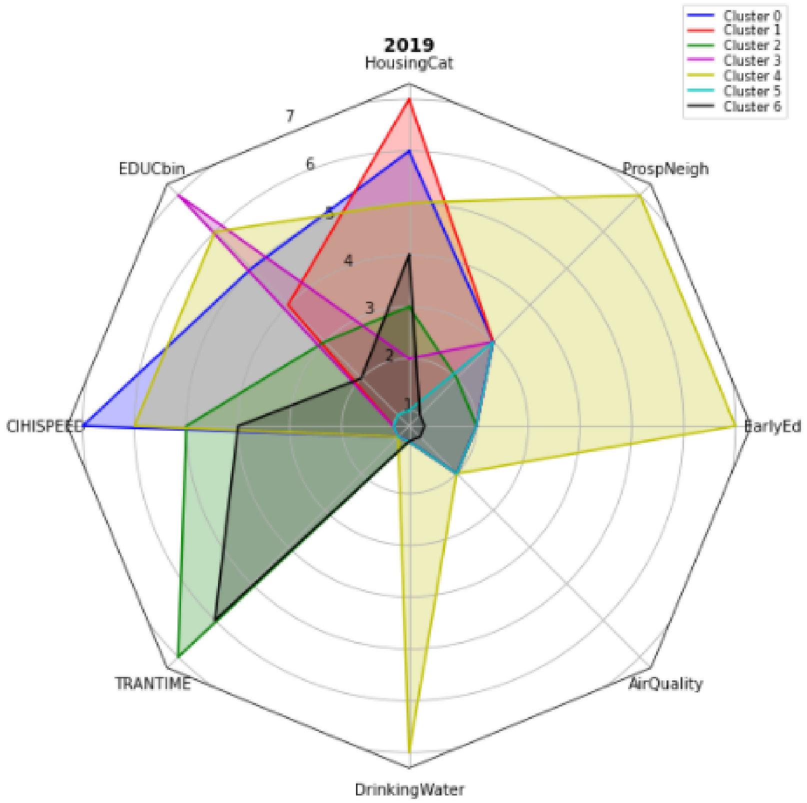
labels	HousingCat	EDUCbin	CIHISPEED	TRANTIME	DrinkingWater	AirQuality	EarlyEd	ProspNeigh	HHINCOME	HISPAN
0	1.335146	0.312078	20.000000	14.168001	0.998420	35.990576	0.509800	0.807171	93838.966758	0.622599
1	0.000000	0.327705	10.000000	13.590036	0.998949	36.728475	0.515279	0.807001	43304.250911	0.614850
2	1.644981	0.523316	10.613623	84.505099	0.997525	36.648086	0.506324	0.813083	153613.361479	0.538017
3	1.773247	0.000000	10.000000	11.703027	0.998370	37.659320	0.492169	0.793149	112202.688141	0.607098
4	1.536120	0.306702	12.456830	12.724906	0.913946	36.858959	0.379997	0.626347	95667.493610	0.498798
5	1.803420	1.000000	10.000000	14.054010	0.998923	37.400179	0.506441	0.801554	166618.212767	0.263557
6	1.642037	0.668558	10.163844	17.484867	0.999969	26.761025	0.603543	0.940831	188576.072549	0.256620

2019 Comprehensive Cluster Feature and Income Medians:



labels	HousingCat	EDUCbin	CIHISPEED	TRANTIME	DrinkingWater	AirQuality	EarlyEd	ProspNeigh	HHINCOME	HISPAN
0	2.0	0.0	20.0	5.0	0.999908	37.0	0.545	0.759245	69200.0	0.0
1	0.0	0.0	10.0	7.0	0.999710	37.0	0.545	0.759245	39064.0	0.0
2	2.0	1.0	10.0	75.0	0.999622	40.0	0.545	0.782130	125229.0	0.0
3	2.0	0.0	10.0	5.0	0.999622	40.0	0.526	0.759245	92687.0	0.0
4	2.0	0.0	10.0	5.0	0.916609	38.0	0.336	0.548841	73304.0	0.0
5	2.0	1.0	10.0	10.0	0.999930	37.0	0.545	0.759245	131032.0	0.0
6	2.0	1.0	10.0	15.0	1.000000	27.0	0.607	0.958168	146758.0	0.0

2019 Comprehensive Cluster Radar Chart:



7.5 Multi-Year Clustering Results

1990 Multi-Year Cluster Category Rankings:

	Housing	*Education*	*Commute*	*HH Income*
Cluster 0	1	4	1	3
Cluster 1	3	1	3	1
Cluster 2	4	3	2	4
Cluster 3	2	2	4	2

2000 Multi-Year Cluster Category Rankings:

	Housing	*Education*	*Commute*	*HH Income*

Cluster 0	1	4	1	3
Cluster 1	2	1	3	1
Cluster 2	4	3	2	4
Cluster 3	3	2	4	2

<u>2010 Multi-Year Cluster Category Rankings:</u>

	Housing	*Education*	*Commute*	*HH Income*
Cluster 0	1	4	1	3
Cluster 1	2	1	3	1
Cluster 2	4	3	2	4
Cluster 3	3	2	4	2

<u>2019 Multi-Year Cluster Category Rankings:</u>

	Housing	*Education*	*Commute*	*HH Income*
Cluster 0	1	1	3	1
Cluster 1	4	3	1	4
Cluster 2	3	2	4	2
Cluster 3	2	4	1	3

<u>Multi-Year Cluster Summaries:</u>

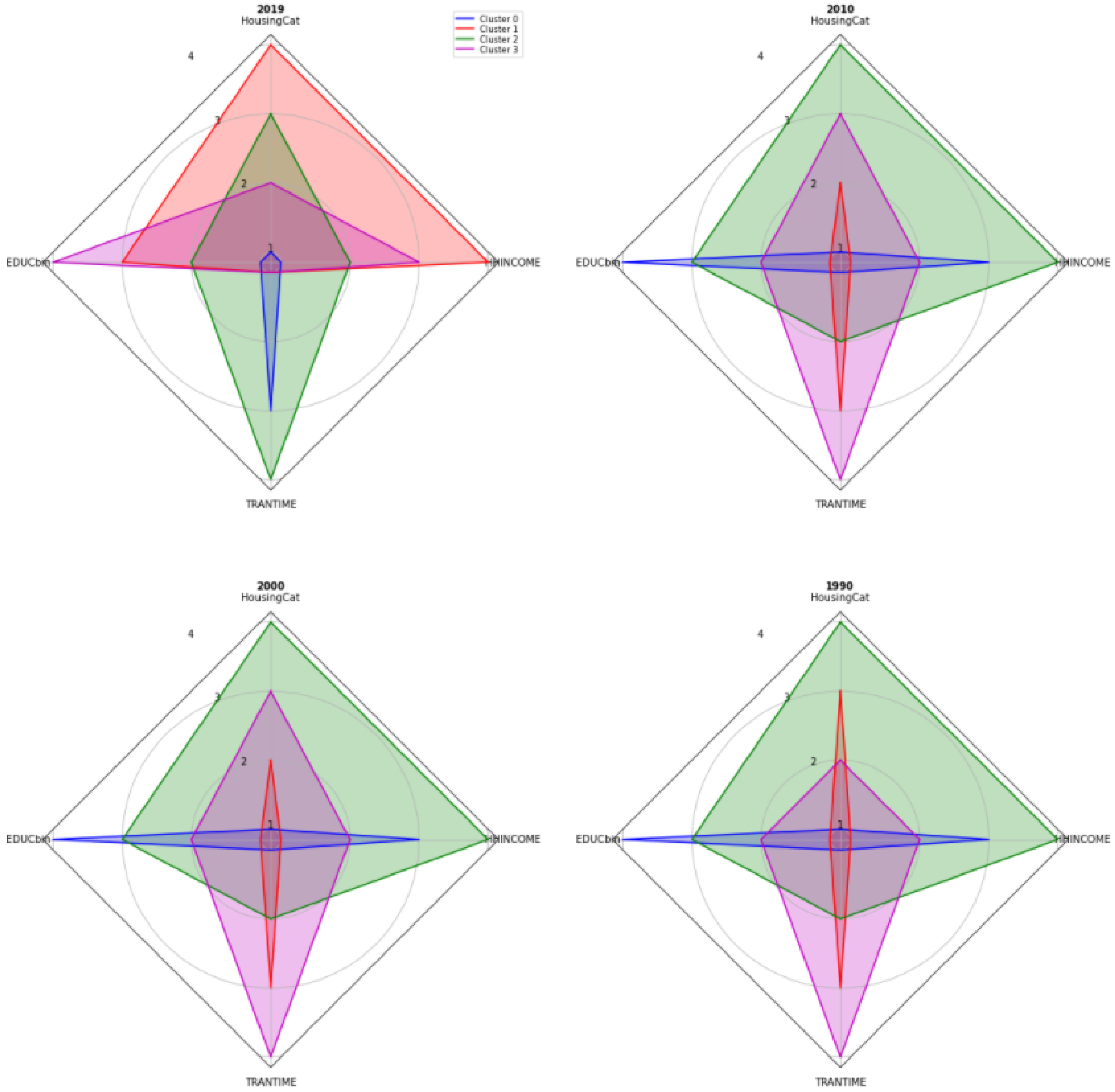
2019
Cluster 0: large group of home-owning, well-educated White and Asian-dominant upper middle class; highest income cluster
Cluster 1: medium group of housing-burdened, less educated Black and Hispanic-dominant lower class; lowest income cluster
Cluster 2: small group of hard working, less educated racially mixed middle class
Cluster 3: large group of less educated Hispanic and Native American-heavy lower middle class

2010
Cluster 0: large group of home-owning, less educated racially mixed lower middle class
Cluster 1: medium group of well educated, White and Asian-dominant upper middle class
Cluster 2: medium group of housing-burdened, less educated Black, Hispanic and Native American-heavy lower class; lowest income cluster
Cluster 3: small group of hard working racially diverse upper middle class; highest income cluster

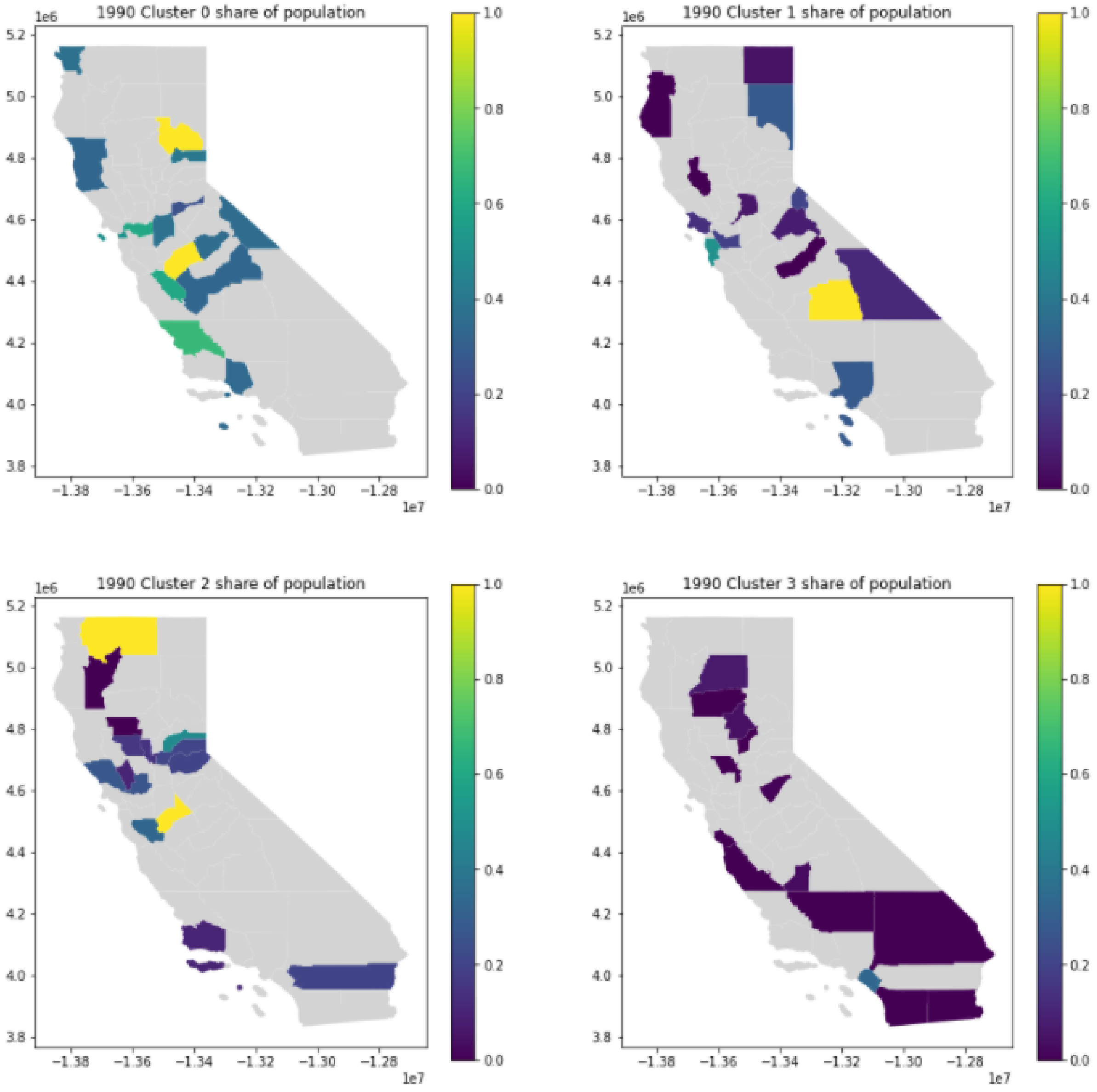
2000
Cluster 0: large group of home-owning, less educated racially mixed lower middle class
Cluster 1: medium group of well-educated, White and Asian-dominant upper middle class; highest income cluster
Cluster 2: large group of housing-burdened, less educated Black, Hispanic and Native-heavy lower class; lowest income cluster
Cluster 3: small group of hard working racially mixed middle class

1990
Cluster 0: large group of home-owning, less educated racially mixed lower middle class
Cluster 1: medium group of well-educated, White and Asian-dominant upper middle class; highest income cluster
Cluster 2: large group of housing-burdened Black, Hispanic and Native American-heavy lower class; lowest income cluster
Cluster 3: small group of hard working racially mixed middle class

Multi-Year Cluster Radar Charts:



Multi-Year Cluster Geographic Makeups:



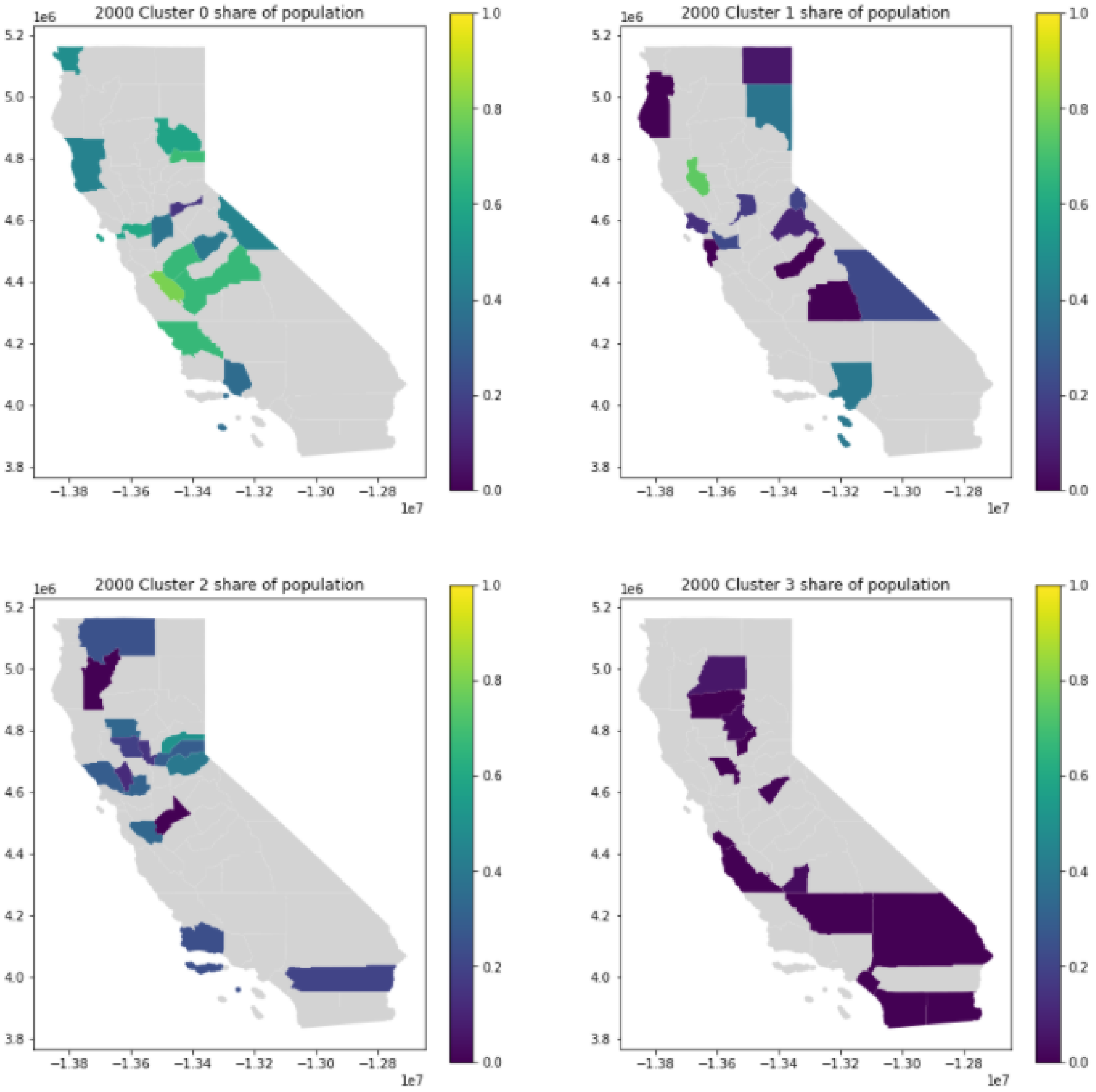
1990 Cluster 0 share of population

1990 Cluster 1 share of population

1990 Cluster 2 share of population

1990 Cluster 3 share of population

1990



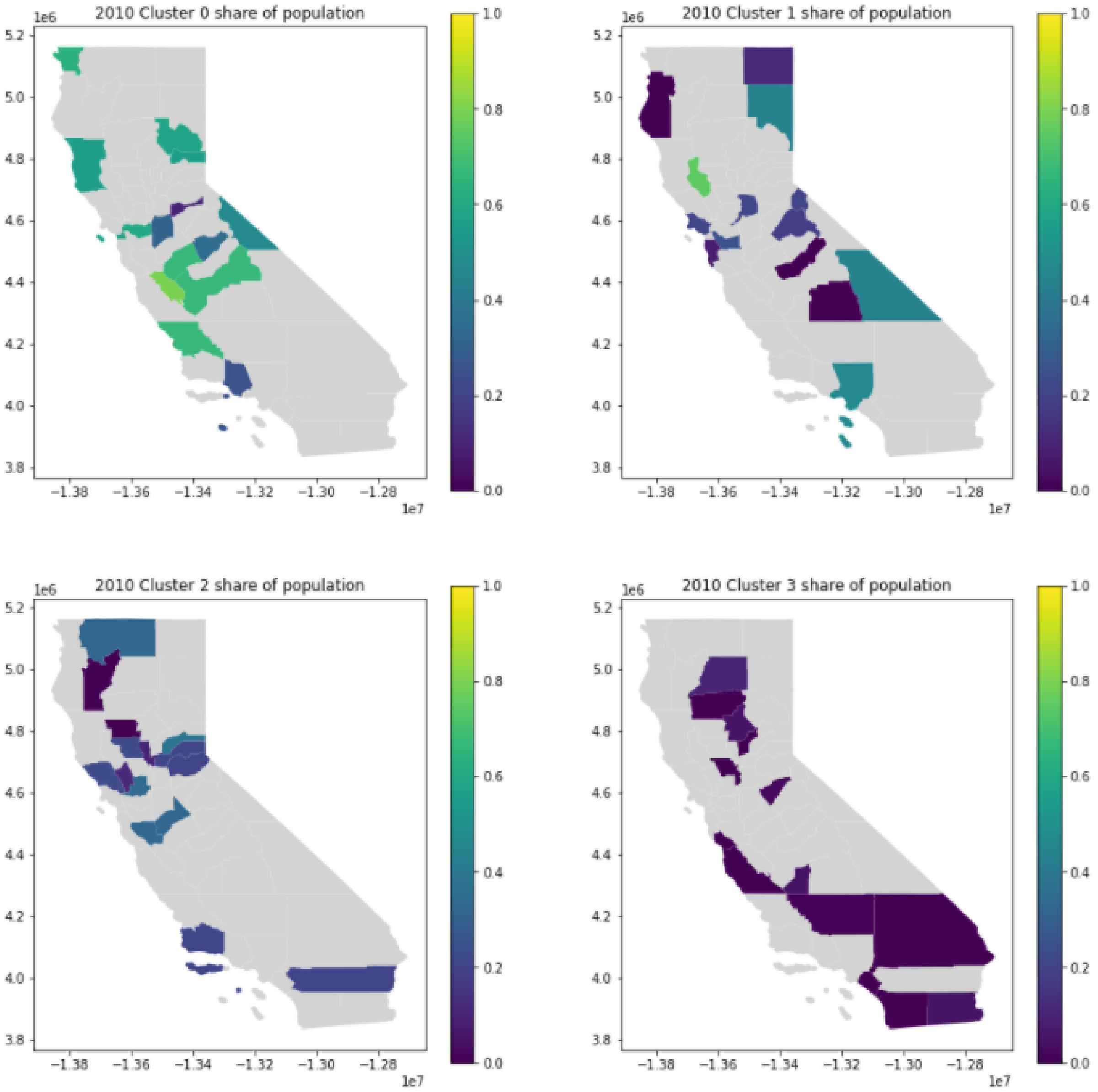
2000 Cluster 0 share of population

2000 Cluster 1 share of population

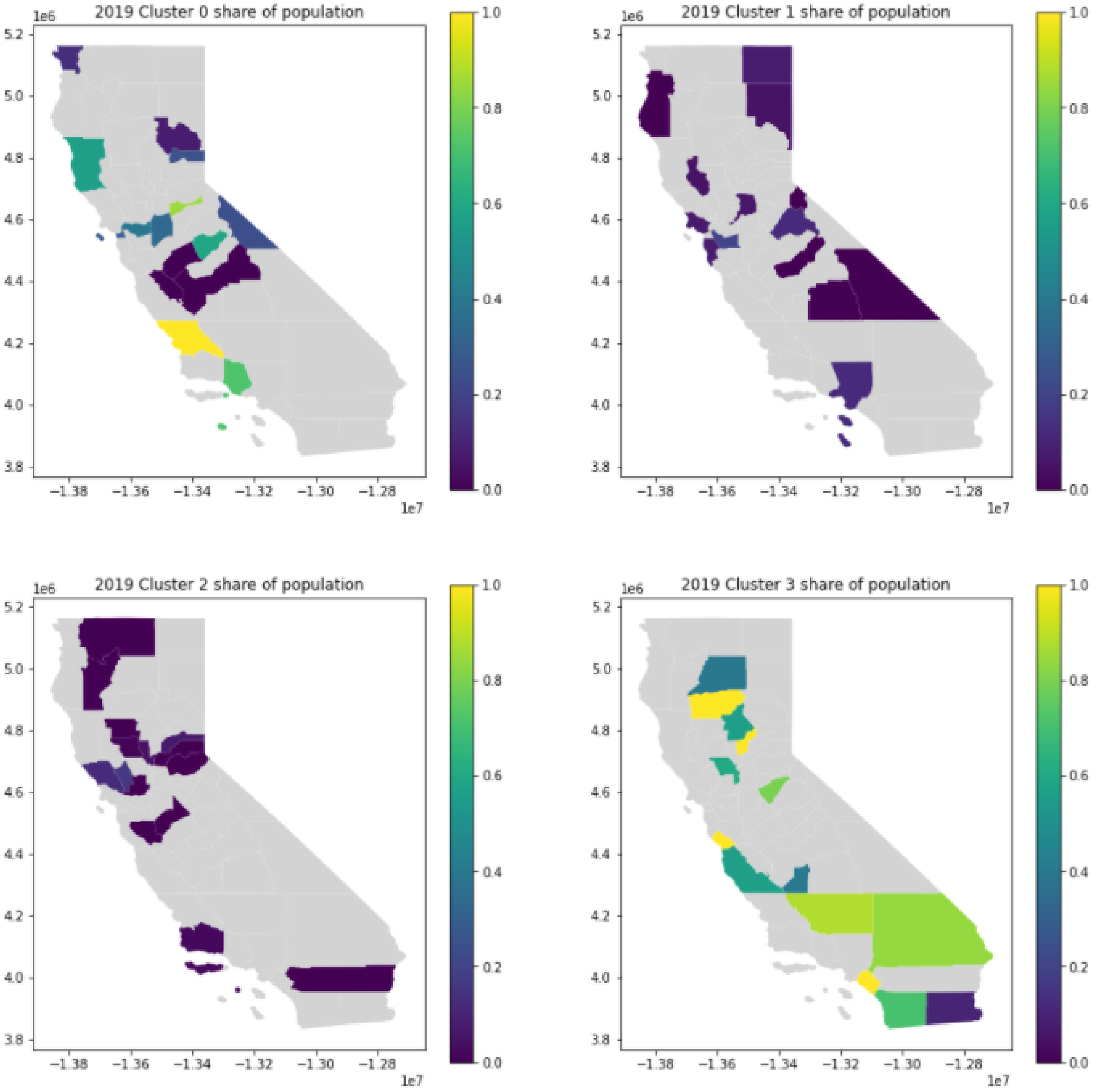
2000 Cluster 2 share of population

2000 Cluster 3 share of population

2000


2010 Cluster 0 share of population
2010 Cluster 1 share of population
2010 Cluster 2 share of population
2010 Cluster 3 share of population
1e6
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
−1.38
−1.36
−1.34
−1.32
−1.30
−1.28
1e7
1.0
0.8
0.6
0.4
0.2
0.0

2010


2019 Cluster 0 share of population
2019 Cluster 1 share of population
2019 Cluster 2 share of population
2019 Cluster 3 share of population
1e6
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
−1.38
−1.36
−1.34
−1.32
−1.30
−1.28
1e7
1.0
0.8
0.6
0.4
0.2
0.0

2019