Classifying Neighborhoods Impacted by Upzoning in California: Methodology

I. Data and Overview

The first step in the process was to identify all of the qualifying transit stations in California. The proposed legislation defines high-quality transit as:

- any kind of fixed rail
- a ferry terminal served by either a bus or rail transit service
- a bus station that has:
    - average headways[1] of 15 minutes or less during the morning peak (6-10am) and evening peak (3-7pm)
    - average headways of 20 minutes or less during weekdays (6am-10pm)
    - average headways of 30 minutes or less on weekends (8am-10pm)

We downloaded General Transit Feed Specification (GTFS) data from over 100 transit agencies in California and created an algorithm to calculate headways for each of the service periods. This work identified 10,550 transit stations that qualified under the SB 50 criteria as of October 2018. After generating a ½-mile buffer around rail stations and a ¼-mile buffer around bus stations, we overlaid the census tract geographies to find the census tracts within these station areas.

To understand the different kinds of neighborhoods that would be impacted by SB 50, we identified 23 characteristics that are important determinants of the character of a neighborhood as well as measures that indicate residents' vulnerability to displacement. Many of the characteristics were modeled after the work of Salon (2015).[2] A list of the characteristics can be found in Table 1. We obtained these variables at the census-tract level from the 2013-2017 ACS 5-Year Estimates.

Table 1: Characteristics Used When Determining Neighborhood Typology

Population Characteristics	Economic Characteristics	Built Form Characteristics
• Share of households that rent • Racial breakdown • Poverty status by race • Households below 200% of poverty line • Households with children • Median age of population • Population with bachelor's degree	• Median rent compared to the county median rent • Vacancy rate of housing units • Unemployment rate • Number of jobs within commuting distance[3]	• Share of housing units that are detached single-family homes vs. multifamily • Share of housing units that were built before 1950 and since 2000 • Population density

[1] A headway is how frequently buses arrive at a certain stop. If the headway is 15 minutes, then a bus arrives every 15 minutes.
[2] Salon, Deborah (2015), "Heterogeneity in the relationship between the built environment and driving: Focus on neighborhood type and travel purpose," *Research in Transportation Economics*, 52 (2015), 23–45.
[3] Commuting distance is defined as the median commute distance for all jobs in a region. This typical commuting distance radius is applied around the population-weighted centroid of each census tract to calculate the total number of jobs within commuting distance.

The final dataset consists of 10,550 rail and bus stations with 23 characteristics. To turn this information into a typology of neighborhoods we used a clustering procedure. First, we reduced the dimensionality of the dataset by applying Principal Components Analysis (PCA). The dataset is a great candidate for PCA since there are many observations and variables and a high degree of correlation between many of the demographic variables. When PCA finds the components of greatest variance between the observations, it accounts for this collinearity and emphasizes the most significant variation. We used the top 10 PCA components based on amount of variation explained to cluster the observations in a k-means clustering procedure and specified the algorithm to create 5 clusters.

We modeled the methodology for this research after Ibes (2015).[4] Ibes uses factor analysis and k-means clustering to classify the public park system in Phoenix, Arizona. The dataset in that paper covered 162 public parks with 14 related characteristics. Using PCA, Ibes reduced the number of factors to five components and applied k-means, identifying five different park types. Song & Knaap (2007) follow a similar approach for classifying neighborhoods in the Portland, Oregon metropolitan area.[5] Their dataset consisted of 6,788 parcels with 21 urban form characteristics. Using factor analysis they reduced the dataset to eight factors, applied k-means cluster analysis to the reduced dataset, and found six different clusters.

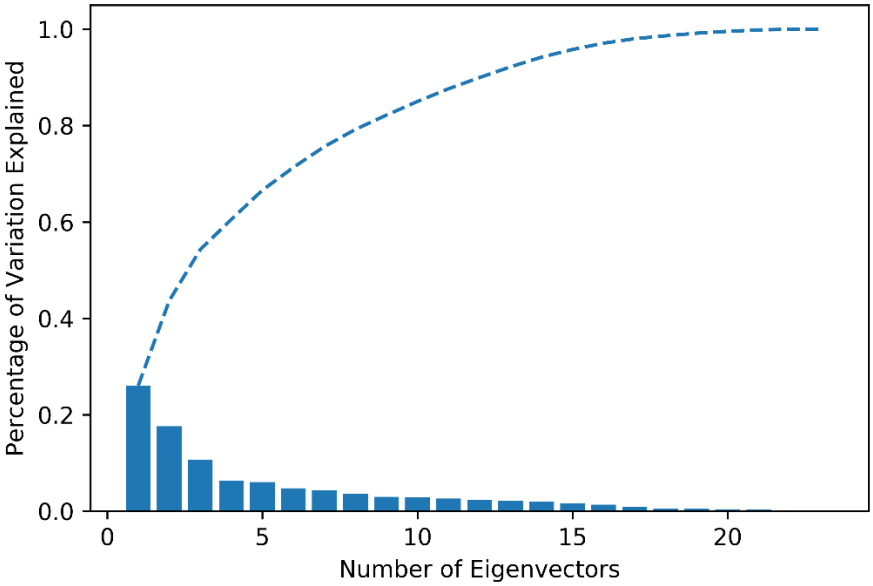
II. Principal Component Analysis

The first step of the analysis is to apply PCA to the dataset. PCA is necessary in this case because many of the neighborhood characteristics in the dataset are highly correlated. For example, poverty and race are often positively correlated. Similarly, income is positively correlated with education. As a result, it is more efficient to collapse these correlated variables into a smaller set of components that removes redundancy and accounts for the correlation. PCA accomplishes this exact task by determining new factors that explain the most variation in the data. In the case, PCA was applied to the dataset using the scikit-learn Python library. Figure 1 shows what percentage of the variation that each resulting factor explains, and the cumulative amount of variation explained.

[4] Ibes, Dorothy (2015), "A multi-dimensional classification and equity analysis of an urban park system: A novel methodology and case study application," *Landscape and Urban Planning*, 137 (2015), 122–137.

[5] Song, Yan and Gerrit-Jan Knaap (2007), "Quantitative classification of neighborhoods: The neighborhoods of new single-family homes in the Portland metropolitan area," *Journal of Urban Design*, 12:1 (2007), 1–24.

Figure 1: Variation Explained by Principal Components



The additional percentage of variance explained decreases rapidly as more factors are added, so only the most important factors should be retained. Ibes selected the top five factors that explained 77.2% of the total variation. Song and Knaap selected the top eight factors that explained 82% of the total variation. In this case, we retain the top ten factors that explain 85% of the total variation in the dataset.

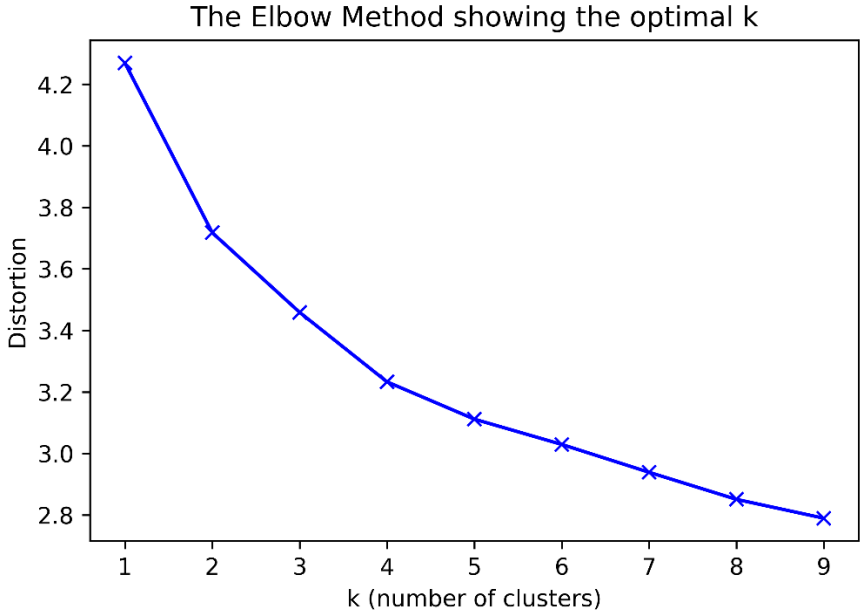
III. Clustering Procedure

The next step in the process is to apply a clustering algorithm to the dataset. Both Ibes and Song and Knapp use the k-means clustering algorithm, which "is a simple unsupervised machine learning algorithm that groups a dataset into a user-specified number (k) of clusters."[6] Since the user specifies the number of clusters, it is necessary to determine what the appropriate number of clusters is. There are a few methods for identifying this parameter. One of the most popular approaches is the "elbow method." In the elbow method, k-means is used to calculate clusters with a range of values for k. For each value of k, the method calculates the "distortion," which is essentially the sum of squared errors between all of the points in a cluster and the cluster centroid.[7] A lower value is better and means the observations are more closely clustered around the centroid. More clusters reduce the error, but there is a tradeoff as increasing the number of clusters past a certain point creates a less meaningful result. To find the best k, look for the "elbow" in the plot, which is where adding one more cluster has the most impact in reducing the distortion.

Figure 2 shows the elbow results for this dataset. Unfortunately, there is no obvious elbow. There are some small kinks at k = 2 and k = 4, but in general there is a smooth decrease in distortion ask increases.

[6] "Using the elbow method to determine the optimal number of clusters for k-means clustering," Robert Gove's Block, https://bl.ocks.org/rpgove/0060ff3b656618e9136b.
[7] "kmeans elbow method," Python, https://pythonprogramminglanguage.com/kmeans-elbow-method/.

Figure 2: Using the Elbow Method to Determine Clusters



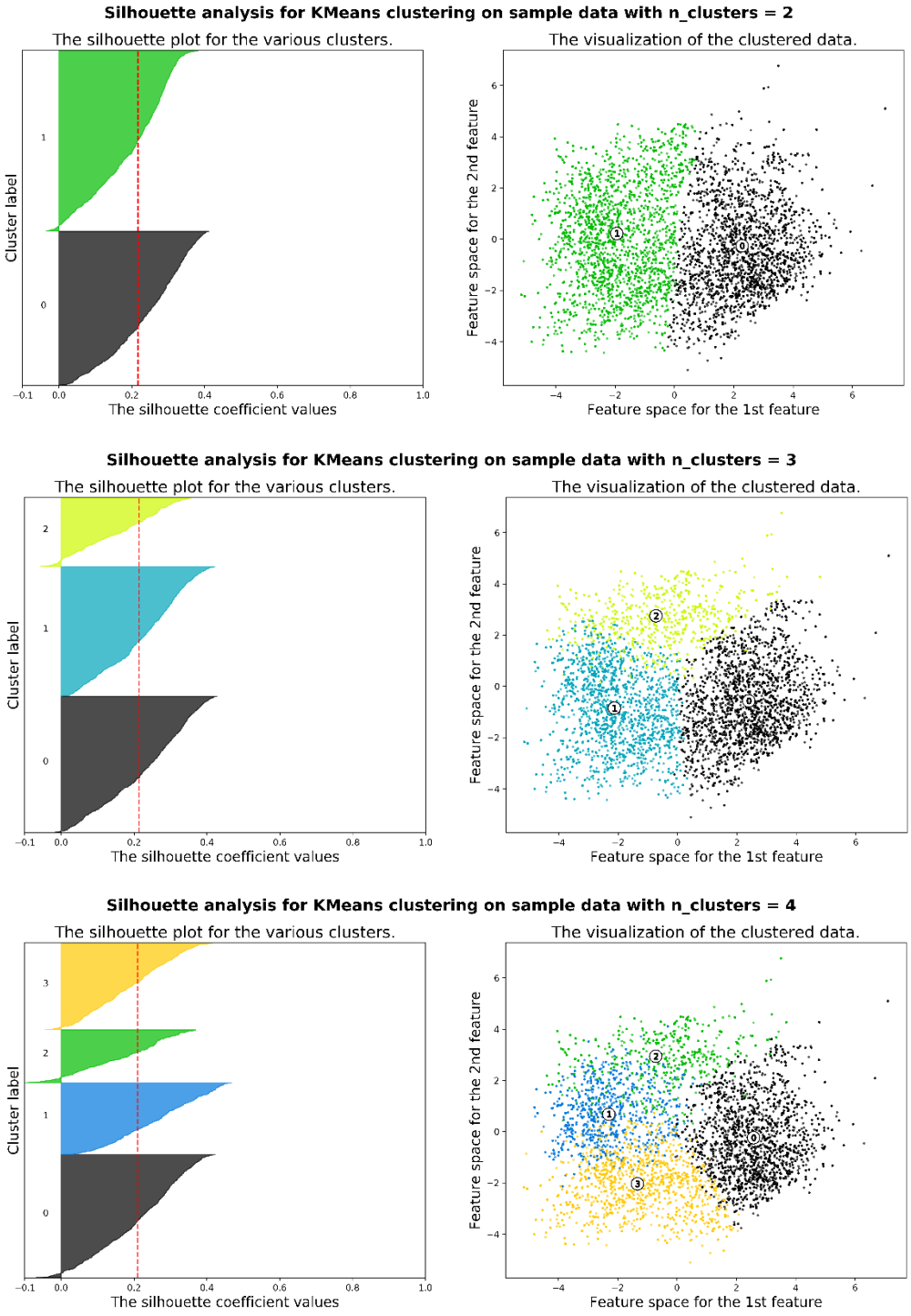
Since the elbow method does not produce a conclusive result, it is necessary to consider another method for identifying the appropriate number of clusters. Figure 3 shows the results for the "Silhouette Method." Rather than studying the compactness within each cluster like the elbow method, silhouette analysis measures the separation of clusters from each other:

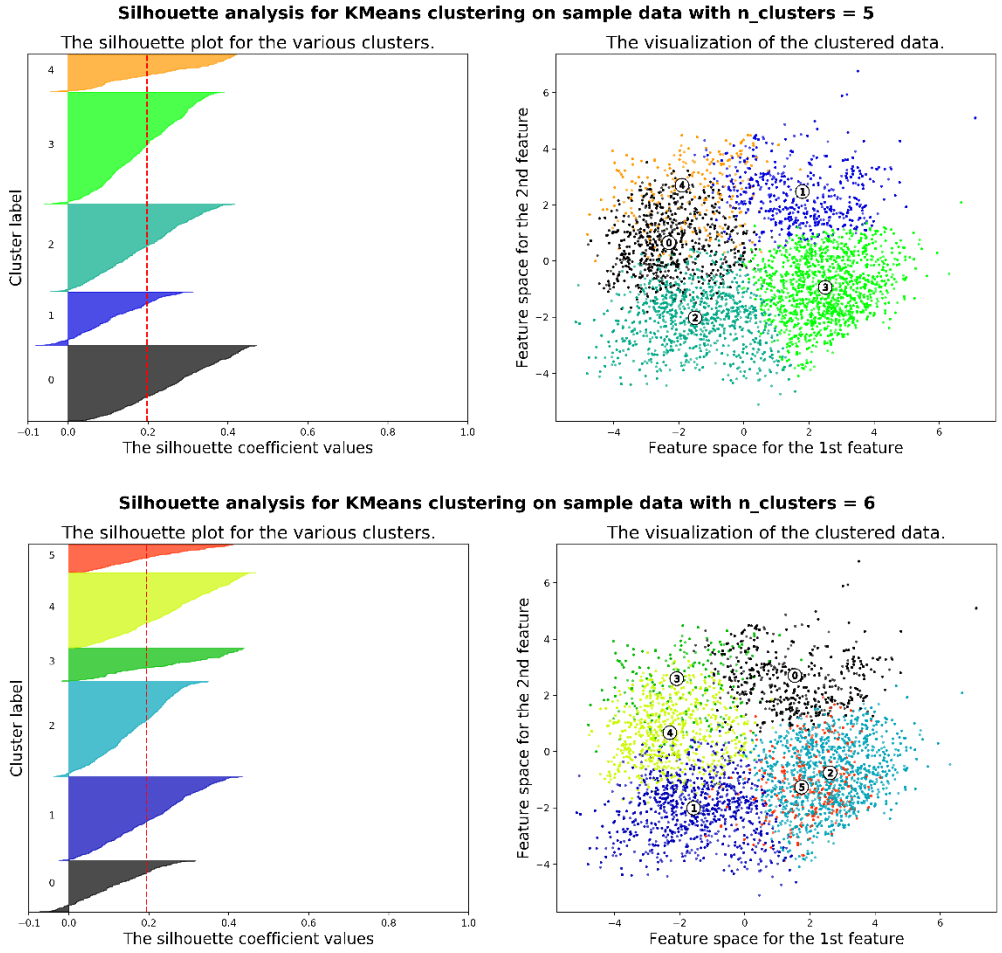
> Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.[8]

Figure 3 shows that the average silhouette coefficient is pretty consistently around 0.2 for values of k from 2 to 6. The iteration with six clusters produces the worst results out of all of the iterations. There is not much separating the other four clustering results, although it appears that k = 3 and k = 4 might be slightly superior from a visual inspection.

[8] "Selecting the number of clusters with silhouette analysis on KMeans clustering," scikit-learn, https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py.

Figure 3: Using Silhouette Analysis to Determine Clusters





Neither of these approaches produces a conclusive result for the number of clusters to choose. Because of this finding, we used the k-means result with k = 5. The results with five clusters make intuitive sense and match our prior understanding of neighborhoods in California.

IV. Cluster Results

The next step is to analyze the cluster results and define the neighborhood typology. After running the clustering algorithm and choosing the appropriate set of results, we returned to the original 23 characteristics and calculated averages for each of the clusters. Table 2 presents these averages.

Table 2: Average Characteristics for Clusters

Cluster	High Density High Income	High Density Low Income	Low Density Low Income	Low Density High Income	Low Density Diverse
Number of stops	963	1,557	3,305	2,186	2,539
Average population	9,231	12,104	10,699	11,692	9,280
Percent of population that rents	74.7%	92.0%	69.6%	71.1%	40.1%
Percent NH White	46.0%	20.7%	7.7%	57.0%	32.9%
Percent Hispanic	16.8%	41.0%	66.8%	14.8%	27.6%
Percent Black	7.6%	9.7%	15.7%	5.1%	7.1%
Percent Asian	25.4%	25.2%	7.3%	17.9%	27.9%
Percent below 200% of poverty rate	31.4%	61.2%	60.4%	24.2%	25.8%
Percent Hispanic in poverty	23.4%	38.0%	29.7%	13.9%	12.7%
Percent Black in poverty	28.1%	44.8%	33.5%	20.8%	15.0%
Percent Asian in poverty	18.7%	30.4%	22.4%	12.9%	9.6%
Percent White in poverty	14.7%	28.5%	25.9%	9.5%	9.1%
Unemployment rate	6.4%	10.8%	11.9%	6.1%	7.4%
Percent with bachelor's degree	60.9%	29.6%	12.2%	62.7%	39.0%
Percent of households with children	12.4%	20.7%	45.9%	16.8%	33.1%
Percent single-family detached house	6.2%	6.5%	41.7%	17.6%	57.2%
Percent small multifamily (2-4 units)	4.1%	8.0%	16.2%	25.6%	8.9%
Percent medium multifamily (5-18 units)	10.2%	22.5%	18.6%	30.9%	8.8%
Percent big multifamily (20+ units)	75.7%	59.2%	12.2%	19.1%	10.1%
Percent of housing units vacant	12.6%	9.0%	5.9%	7.2%	5.1%
Percent of units built before 1950	17.9%	41.4%	40.4%	50.5%	33.4%
Percent of units built after 2000	36.5%	13.1%	5.8%	4.9%	6.0%
Average population/square mile	11,639	26,631	15,634	21,620	11,142
Median tract rent / median county rent	1.32	0.76	0.81	1.14	1.12
Jobs within commuting distance	1,092,714	1,465,269	1,187,058	1,093,013	790,501

Patterns emerge when comparing these averages across clusters. We defined clusters by visually inspecting this table and noting when a cluster had an average that was well above or below the other clusters. Table 3 presents the results of this process. The colors in the column headers match the colors of the circles in the interactive map.

Table 3: Cluster Descriptions

High Density High Income	High Density Low Income	Low Density High Income	Low Density Low Income	Low Density Diverse
Whiter	More People of Color	Whiter	More People of Color	Racially Diverse
Low Poverty	High Poverty	Low Poverty	High Poverty	Low Poverty
Lower Density	High Density	Medium Density	Medium Density	Low Density
Big Multifamily	Big Multifamily	Medium Multifamily	Single-Family	Single-Family
High Cost	Low Cost	High Cost	Low Cost	High Cost
Newer Buildings	Older Buildings	Older Buildings	Older Buildings	Older Buildings
High Education	More Renters	High Education	Low Education	More Owners
	More Job Access			Less Job Access

V. References

Ibes, Dorothy (2015), "A multi-dimensional classification and equity analysis of an urban park system: A novel methodology and case study application," Landscape and Urban Planning, 137 (2015), 122–137.

"kmeans elbow method," Python, https://pythonprogramminglanguage.com/kmeans-elbow-method/.

Salon, Deborah (2015), "Heterogeneity in the relationship between the built environment and driving: Focus on neighborhood type and travel purpose," Research in Transportation Economics, 52 (2015), 23–45.

"Selecting the number of clusters with silhouette analysis on KMeans clustering," scikit-learn, https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py.

Song, Yan and Gerrit-Jan Knaap (2007), "Quantitative classification of neighborhoods: The neighborhoods of new single-family homes in the Portland metropolitan area," Journal of Urban Design, 12:1 (2007), 1–24.

"Using the elbow method to determine the optimal number of clusters for k-means clustering," Robert Gove's Block, https://bl.ocks.org/rpgove/0060ff3b656618e9136b.