ISyE 6740 - Fall 2023

Team 100 Project Report

Author: Charles Kramer ckramer36 GTID 903744461

Virginia: Finding Commonality Amid Division

Problem Statement

It is commonplace to point out divisions in American society. The United States is a large and diverse country, so differences abound. Contrasts, particularly ones that play into readers' existing cognitive distortions or framing, are intellectually and emotionally stimulating, and humans have an intellectual tendency to reduce complexity to simplicity [Thaler, 1994]. Consider, for example, the caricatures of conservative versus liberal, rural versus urban, and South versus North. These frames are useful to promote understanding, but they mask considerable diversity within each of these categories.

U.S. political and social divisions came to the fore of public discussion with the 2016 elections. Indeed, divisions seem to have become more pronounced with time, fueled by social media. And in the state of Virginia, these divisions have come into sharp focus in the context of the midterm 2023 elections and the consequent battle for control of the state legislature [Yancey, 2023].

The political questions in Virginia have gained national interest. Against the backdrop of national social and political divisions, Virginia has become something of a political bellwether for the nation [Myrick, 2023], [Gabriel, 2023]. Going into the election, Republicans controlled the state House of Delegates, Democrats controlled the state Senate, all seats were open for election, and redistricting had reshaped many districts. In this context, control of the state legislature was in play. And accordingly, crucial policy questions that align with political and social divisions–abortion, gun control, education–were also at stake. As a result, donors and Political Action Committees poured money into Virginia state races. Even local elections for positions such as school board or county supervisor were hotly contested and attracted unusual levels of attention. They also attracted record levels of spending, with media outreach and policy positions often playing to divisions around national themes.

And consistent with the human inclination to simplify, the public discussion around Virginia state elections often draws sharp distinctions between 'liberal, urban' Northern Virginia and the 'conservative, rural' West and South. Indeed, liberal donors mostly declined to put resources into races in Southwest Virginia, regarding them as safely in conservative hands. Similarly, conservative parties have struggled to achieve a foothold in some of Virginia's Northern population centers. These trends reinforce the caricatures of rural versus urban, conservative versus liberal.

But this caricature neglects important regional nuances. Quite a few of Virginia's major universities–including top ranked University of Virginia and Virginia Tech–are based well outside Northern Virginia. With the advance of technology, high-speed internet access is improving statewide. Virginia is also known for its data centers–the state is home to 35 percent of hyperscale data centers worldwide–and while these are concentrated in the Northeast, they can be found in the South and West as well (See Figure 1). The Defense Department has facilities across the state including the headquarters of the Department of Defense (the Pentagon in Arlington) and the world's largest naval base (Naval Station Norfolk), which along with the surrounding defense industry brings considerable technological expertise to the state.

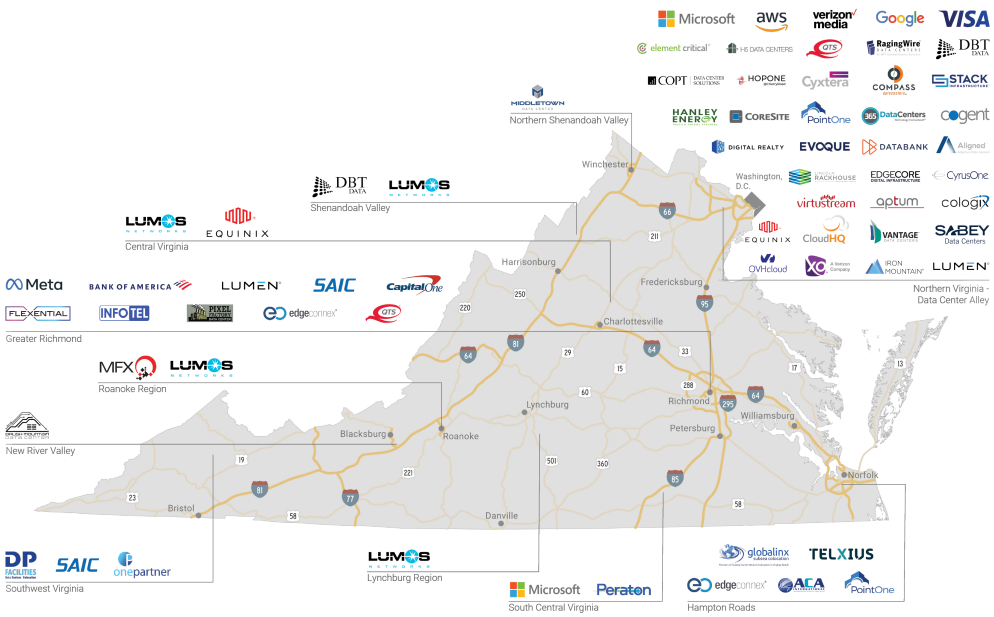


Figure 1: Data Centers in Virginia (source: Virginia Economic Development Partnership)

In addition, geographic areas–even small ones–may not be as homogenous as generally thought. Portraying a particular large geographic area as solely characterized by a particular set of economic, social and demographic characteristics can be misleading. For example, political scientists have identified that people tend to live in clusters of politically homogenous 'bubbles' of people with similar beliefs. But at the same time, there can be nearby clusters of people with different beliefs [Wezerek et al., 2021]. Similarly, research on political geography shows that while people of similar political stripes tend to cluster together, there can be considerable diversity even at the district level [Rodden, 2019], [McCarty et al., 2018].

Accordingly, I attempt to develop a more nuanced understanding of the social, economic and demographic map of Virginia. I search for pockets of commonality across regions of Virginia, bearing in mind that these pockets may be geographically small. I apply clustering techniques to data at the census-tract level in search of such pockets. I examine social and economic characteristics for the clusters and look for 'outliers' that are geographically far from clusters that have similar characteristics, in search of commonalities across regions.

Data Sources

I draw data from two main sources: the US Census and the Yelp API.

For Census data, I employ the American Community Survey (ACS) 5-year estimates for 2021. I drew 208 initial variables from the ACS, collectively characterizing demographics (age, sex, race), languages spoken, industry of employment, tenure in present location, rental vs ownership, availability of computers and internet access in the home, income (including social assistance), health insurance status, and relationship status (married, divorced, cohabiting, etc). These data are extracted at the census-tract level, a construct of variable geographic area but meant to represent an average of 4000 persons with somewhat uniform characteristics [Census,]. There are a total of 2098 census tracts in Virginia employed in the ACS. See Table 1 for depictions of census tracts in Arlington County (in Northern Virginia, on the border with the District of Colombia) and Patrick County (in Southern Virginia, on the border with North Carolina); Table 2 presents selected data for the two sample census tracts depicted in Table 1. I accessed the tract-level data via the Census API, using the 'census' package in Python. After dropping census tracts that have no land or are missing values, the dataset had 178 variables and 2168 tracts, for about 385,000 datapoints. I also extracted latitude and longitude for each tract from the Census Gazeteer (to be used in constructing the Yelp database).

I reduced the dimension of the ACS data using principal components (PC) analysis. I selected the number of PCs based on the variance ratio (share of variance explained by the first M PCs) (see Figure 2). The first 5 PCs explain about 60 percent of the covariance across the ACS data, and the 6th and additional PCs explain a small (single digit) and declining share of the covariance.

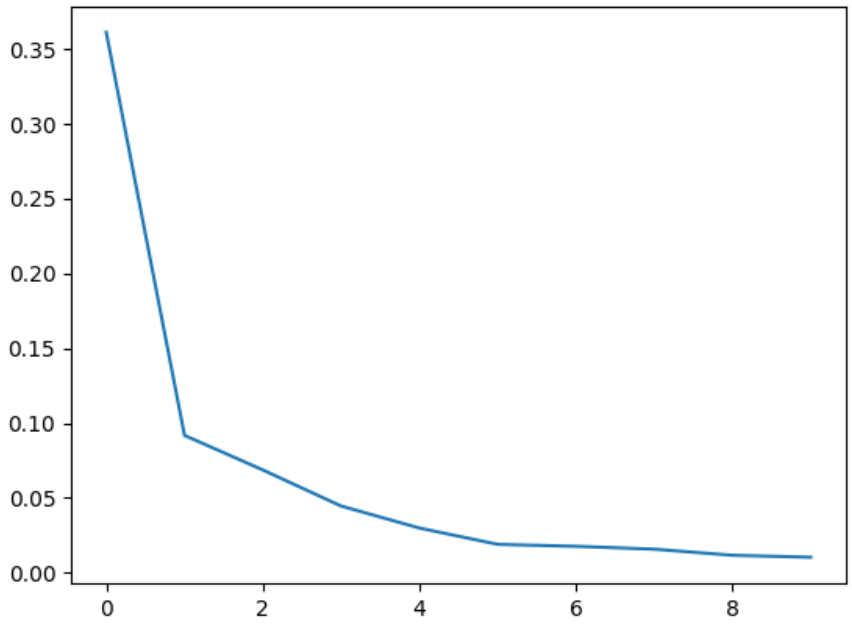


Figure 2: ACS Data: PCA Variance Ratio vs PC Number

The principal components, which are weighted averages of the underlying data, can tell us something about the most important themes in the census data. Examining the largest weights in absolute value for the first three PCs (see Figure 3), according to the broad concept for the respective variable, the PCs have substantial weight on place of birth (black bars), computer and internet use (green bars), housing characteristics (blue bars), school enrollment (light blue bars), relationship to householder (red bar), household type (orange bars), and income (purple bars). For example, the largest weights in the first PC are on a variable representing place of birth, a variable representing the relationship to the householder, and two variables concerning computer and internet use. The differences in weights per category across the PCs show that the PCs measure distinct concepts. For example, the third PC is clearly dominated by housing characteristics, which receive less weight in the other two PCs shown. The categories also provide a sense of what socioeconomic characteristics may be most important in defining clusters.

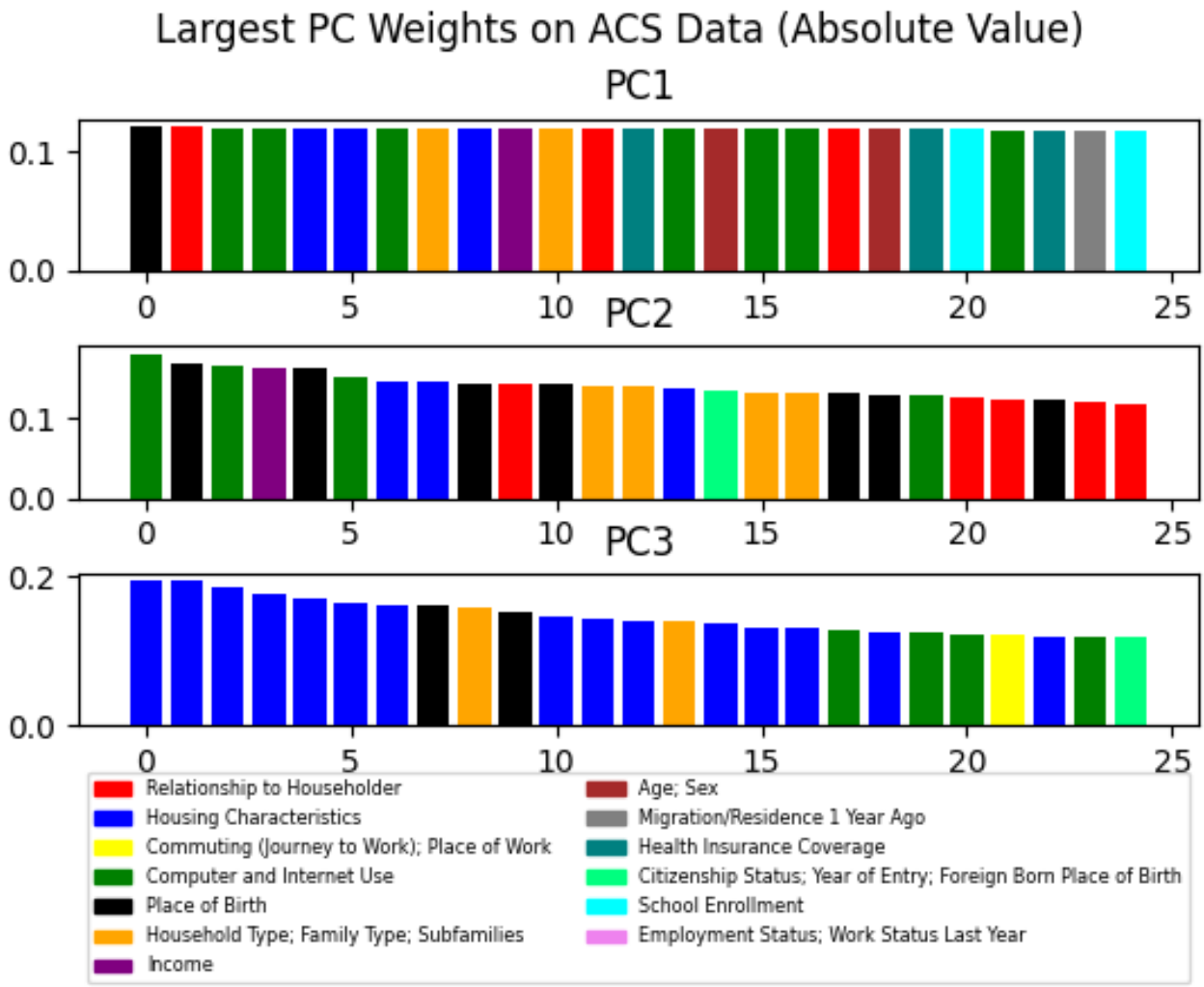


Figure 3: ACS Data: PCA Components by Category

I also extracted Yelp data to characterize the social and political leanings of census tracts [Yelp,]. Researchers have found interesting social patterns using these data. For example, [Wasserman, 2014] performs a simple analysis of partisan leanings and electoral prospects using the number of Whole Foods stores versus the number of Cracker Barrel restaurants. [Caren,] measures local culture using the local predominance of sushi restaurants versus all-you-can-eat buffets. Accordingly, I ran the following process for each census tract in Virginia. I extracted the top five reviews for locations in or near the tract under the keywords 'liked_by_vegetarians' and 'gender_neutral_restrooms' from the Yelp API using the 'requests' package in Python. For each of these two concepts, I then computed the average distance to the reviewed businesses for each keyword. If there were no reviews under the keyword for a census tract, I assigned the tract the maximum distance for that keyword across all tracts.

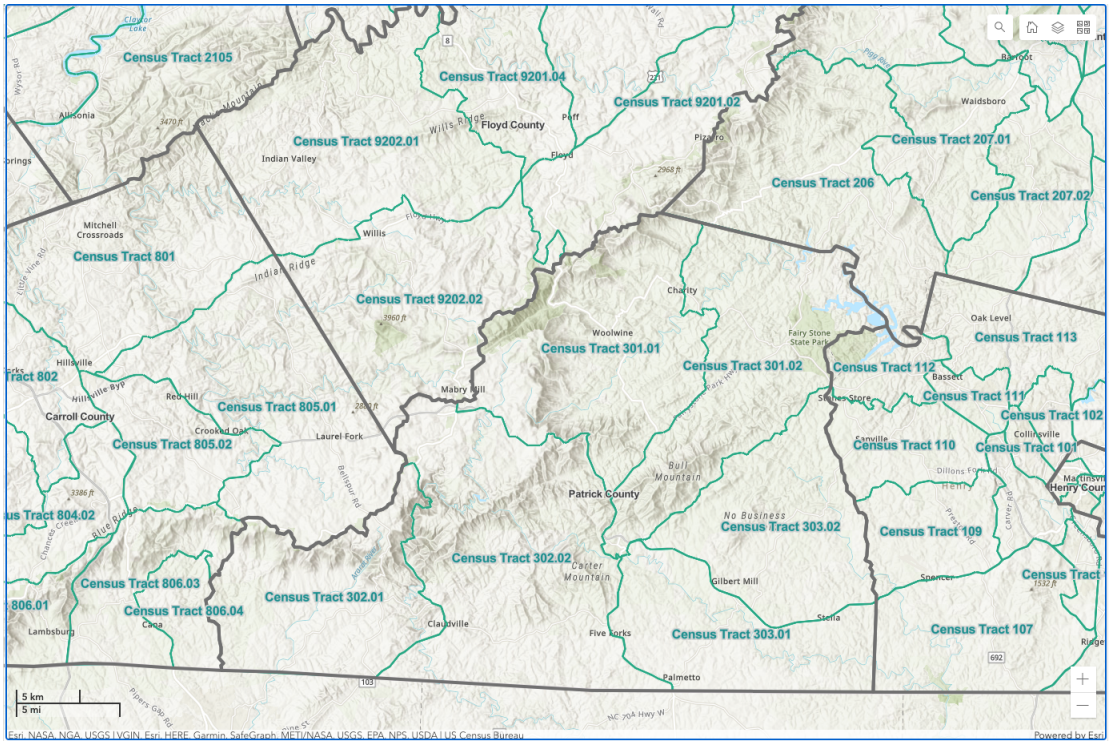


Table 1: Census Tracts: Arlington and Patrick Counties. Note the difference in scale.

	Tract 1019	Percentage	Tract 301.02	Percentage
All households	1,267		922	
With earnings	1,094	86	539	58
With wages or salary income	1,038	82	537	58
With self-employment income	208	16	35	4
With interest, dividends, or net rental income	669	53	123	13
With Social Security income	259	20	514	56
With Supplemental Security Income (SSI)	0	0	29	3
With cash public assistance income or Food Stamps/SNAP	19	1	137	15
With cash public assistance	11	1	31	3
With retirement income	263	21	240	26
With other types of income	72	6	92	10
Total Population:	3,183		2,288	
Population of one race:	2,858	90	2,197	96
White alone	2,488	78	2,007	88
Black or African American alone	102	3	129	6
American Indian and Alaska Native alone	5	0	5	0
Asian alone	225	7	7	0
Native Hawaiian and Other Pacific Islander alone	2	0	2	0
Some Other Race alone	36	1	47	2
Median age (years)	37.0	—	55.2	—
Sex ratio (males per 100 females)	81.5	—	91.4	—
Age dependency ratio	51.3	—	70.8	—
Old-age dependency ratio	25.6	—	46.6	—
Child dependency ratio	25.8	—	24.2	—

Table 2: Selected Census Data: Census Tracts 1019 (Arlington County) and 301.02 (Patrick County)

Clustering Analysis

I used K-means to perform the clustering, first standardizing the data to put them on a common scale. I used the sum of squared errors (squared Euclidean distance to closest centroid) and silhouette score (measure of the precision of classification) to choose K (the number of clusters). The 'elbow plot' for K reveals that the SSE drops continuously for each value of K, without the characteristic flattening of the curve ('elbow') (See Figure 4). However, the silhouette score stabilizes around K=5, a point at which the SSE is reduced by about 30 percent compared with k=2 clusters. I took this to represent a reasonable tradeoff between within-cluster spread and across-cluster differentiation.

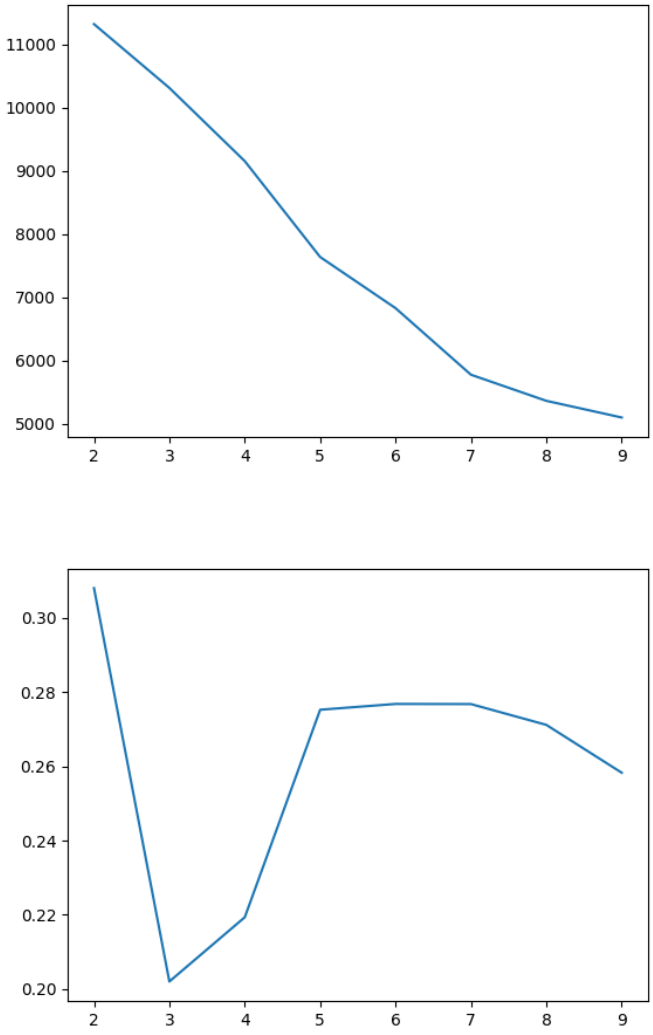


Figure 4: SSE and Silhouette Score vs Number of Clusters (k)

Now plotting the clusters across Virginia (Figure 5), we can see a substantial number of Cluster 3 around Northern Virginia near DC, with Clusters 2 and 4 dominating in the Southwest and Clusters 3 and 4 dominant around Central Virginia (Richmond) and the Atlantic Coast (Virginia Beach/Norfolk). However, it is interesting to note that Northern Virginia has a substantial number of Cluster 2 while Cluster 3 can also be found in the Southwest. Overall, there is no obvious sharp division in regions by clusters.

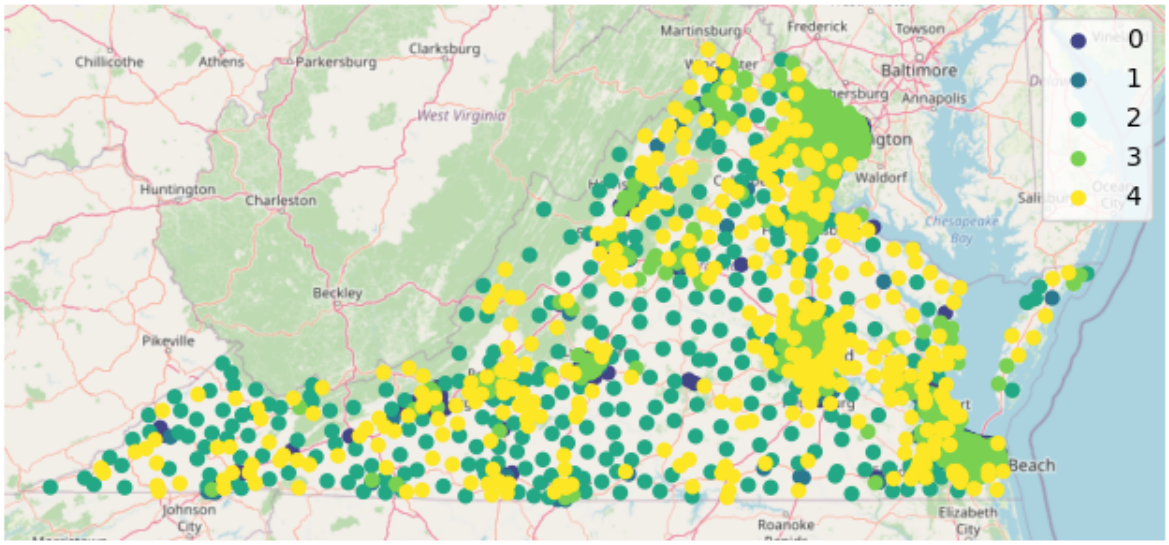


Figure 5: Cluster Map

To gauge the significance of the Yelp data versus Census ACS data, I ran two separate estimations (details not shown) for the Yelp variables alone and ACS variables alone; they show a similar geographic diversity of classification. A plot of the geographic centroids (medians) for each sub-estimation is suggestive of some socio-economic structure in Virginia running North-South versus East-West. Figure 6 shows that the centroids for the combined dataset (in yellow) form a reversed 'L" shape with a North-South branch and an East-West branch. The centroids for the Yelp data alone (red) run broadly East-West, running from about Mechanicsville to the Lynchburg-Appomattox area (one point is orange where it overlaps with the combined dataset), while the ACS-only centroids run North-South from Fredericksburg down to southern suburban Richmond. Accordingly, the Yelp variable–measuring 'culture'–distinguishes between East and West locations, while the ACS variable–'socioeconomics'–distinguishes North from South. The variables draw this distinction from data that have no explicit geographic component. It is also notable that the centroids are distributed near the population center of Virginia (in Goochland County, northwest of Richmond); this is further evidence that the procedure does not result in clusters that are geographically isolated, although there are definite tendencies for certain clusters to predominate in particular regions.

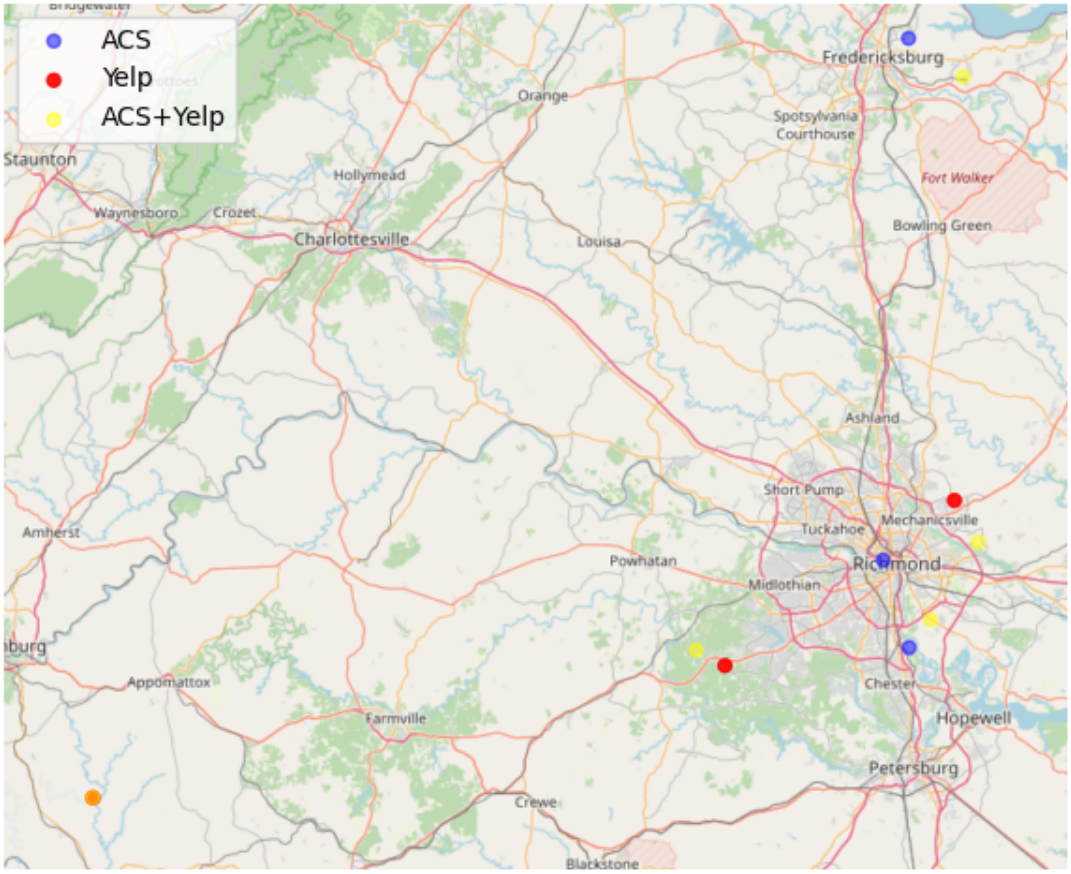


Figure 6: Geographic Centers for Clusters (ACS ('socioeconomic'), Yelp ('cultural'), and Combined Model)

Regional Distribution of Clusters

To make this picture clearer, Figure 7 uses the demographic regions defined by the University of Virginia Weldon Cooper Center for Public Service [WeldonCooper and UVa.,], which they use for analysis of social, demographic and economic trends and patterns in Virginia. This figure maps the distribution of clusters for each of the eight regions (Southwest, West Central, Valley, Southside, Hampton Roads, Eastern, Northern, and Central). Note the stark differences in the distribution of clusters across regions–in Southwest Virginia, Cluster 2 dominates, followed by Cluster 4, whereas Northern Virginia is almost exclusively Cluster 3. Eastern Virginia is heavily dominated by Cluster 4, and Hampton Roads has a distribution closer to Northern Virginia (albeit not as concentrated on one cluster).

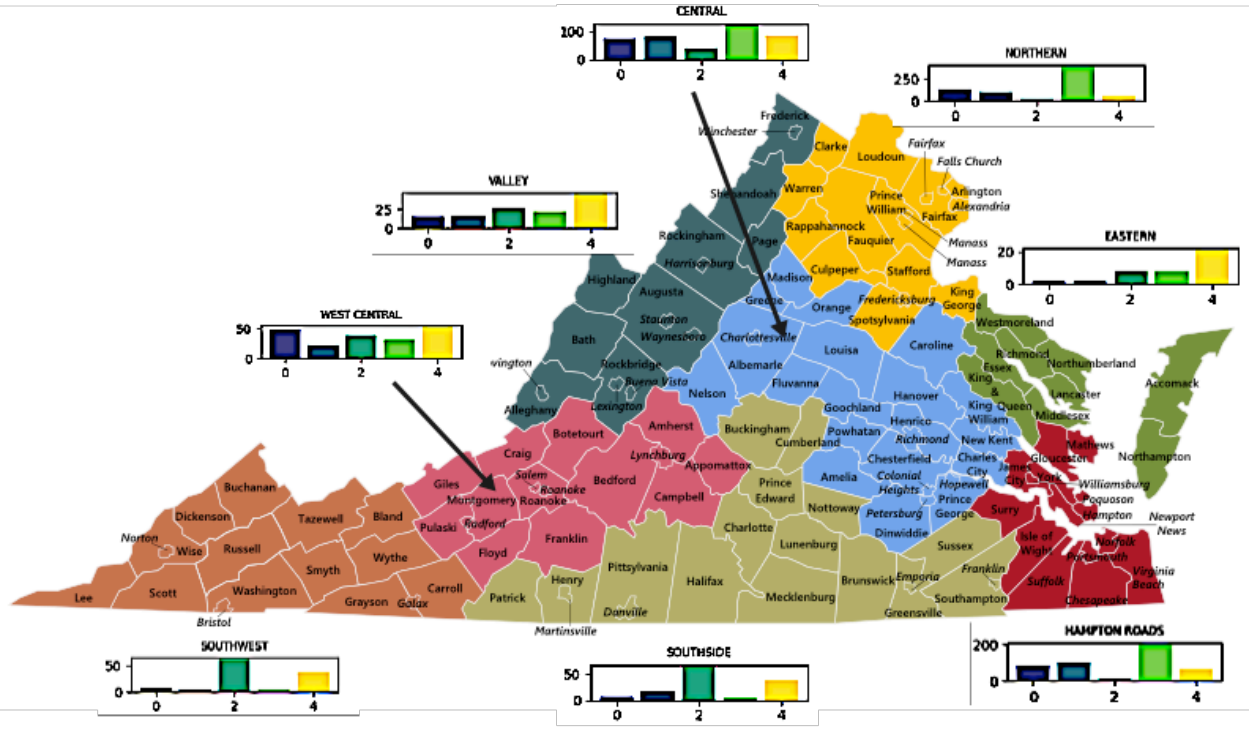
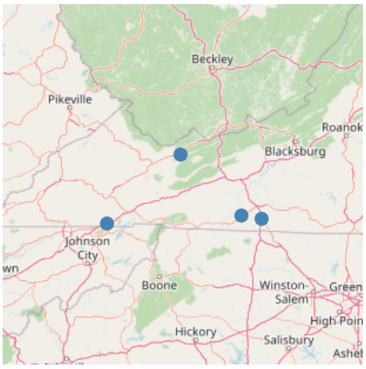
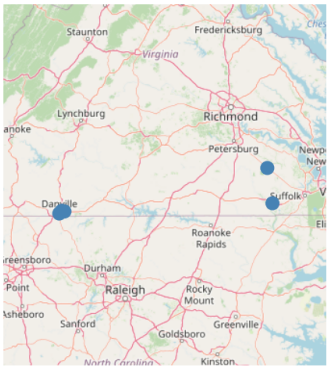


Figure 7: Distribution of Clusters by Region

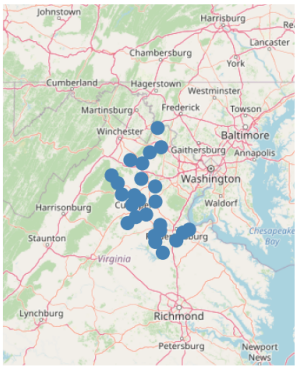
That said, the figures also show 'outliers' for each region–e.g. clusters that are more characteristic of other regions. For example, Southwest and Southside have a handful of Cluster 3 tracts, which are more characteristic of Northern Virginia; similarly, Northern Virginia has a number of Cluster 2 tracts, which dominate in the Southwest. Figure 8 shows these outliers. Panel (a) shows Cluster 3 in Southwest, concentrated in the area between Virginia Tech and Radford University; accordingly it makes sense that these tracts might more closely resemble Northern Virginia. Panel (b) shows Cluster 3 in Southside, with one in Danville on the North Carolina border and two near the border with Hampton Roads. Panel (c) shows the distribution of Cluster 2 in Northern Virginia, which is concentrated along a North-South corridor of exurbs from Winchester to Fredericksburg, framed by Interstate Routes 66, 81, and 95. These figures demonstrate that the distribution of demographic, economic and social characteristics is not uniform within regions; there are tracts in Northern Virginia that most resemble Southwest Virginia, and vice versa.



(a) Cluster 3 in Southwest



(b) Cluster 3 in Southside



(c) Cluster 2 in Northern

Figure 8: Examples of Outlier Clusters

To dig deeper into the differences across clusters, as well as to assess the ability of the K-means

procedure to distinguish the distribution of ACS and Yelp variables across the identified clusters, I ran an Anderson-Darling multisample test for equality of distributions across each variable [Scholz and Stephens, 1987]. That is, for each unprocessed variable in the ACS and Yelp dataset (without standardization or dimensionality reduction via PC) I tested whether its distributions for Cluster 0, Cluster 1, Cluster 2, Cluster 3 and Cluster 4 were identical. The p-values for these tests are shown in Figure 9, including those for Yelp-only and ACS-only runs (p-values are sorted lowest to highest for each run). Note that the P-value is censored at 0.01 and .25 (see [Scholz and Stephens, 1987]). The figure shows that the ACS data do a better job than the Yelp data of creating clusters that have distinct distributions of characteristics; for almost all the variables, we reject the null at the 0.1 percent (one tenth of one percent) level.

To test whether the results are driven by population differences, I selected a handful of variables (those used in the comparison charts shown later) and re-ran the tests using the distribution of the variable as a share of the population in each tract. The results were the same–all rejected the null at 0.1 percent.

Given the number of tests performed, the usual caveat about simultaneous testing applies here. The Bonferroni correction applied to the test results would dictate a conservative P-value of p/n = .1/178 = 0.05 percent at the nominal 10 percent significance level. Because the K-sample critical values are censored at 0.1 percent, it is impossible to discern how many tests meet that standard, although the statistical procedure produced many results where the p-value would have been below 0.1 percent save for the censoring.

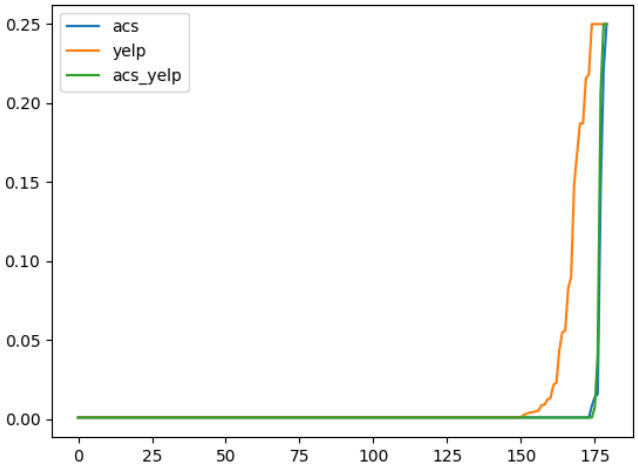


Figure 9: P-values for Tests of Equal Distributions Across Clusters for Each Variable (Sorted)

I next plotted the distribution of selected variables across clusters to look for patterns. I examine 8 variables: the number of persons (i) by population, (ii) in households headed by same-sex spouses, (iii) with income over $75,000, (iv) living below poverty level, (v) living in owner-occupied housing, (vi) working full time without health insurance coverage, (vii) covered by Medicaid, and (viii) with a broadband internet connection.

Figure 10 shows the distribution of population per cluster. Cluster 1 stands out as a more populous, right-skewed cluster, particularly compared to Clusters 0 and 4. Comparing Clusters 2 and 3, Cluster 3 has a greater concentration of highly-populated clusters; this is sensible as Cluster 3 tends to appear in Northern Virginia, with the dense suburbs closer to the population center of Washington DC.

To check whether population drives the results for the remaining charts, I re-calculated them using data for the variables as a share of population–the results were qualitatively the same. I present them

in absolute (number of persons) as absolute numbers are of interest for a number of applications; for instance, a marketing campaign may not want to target a tract with 20 percent potential customers if the actual number of potential customers is low.

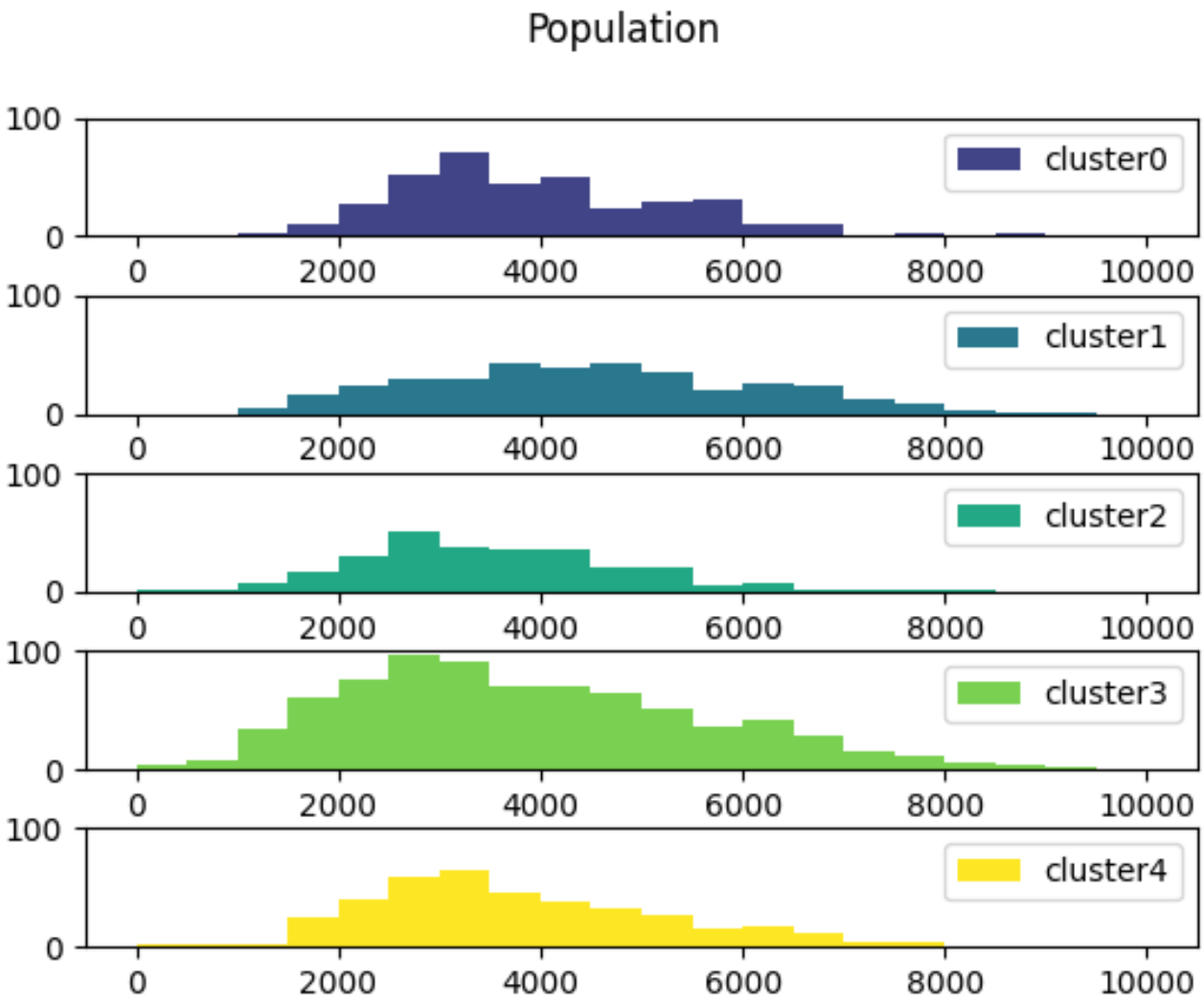


Figure 10: Distribution Across Clusters

Figure 11 shows the distribution of the number of persons in households with a same-sex spouse. The clustering algorithm identifies Cluster 3 as having more persons in same-sex marriage households than Cluster 2, again consistent with the broad rural-urban divide; although the number in same-sex households appear in nontrivial numbers in every cluster. That said, the percentages are small; very few tracts in Cluster 3 have one percent or more of its population same-sex households.

Examining data related to income, Figure 12 shows the distribution of the count of persons with income over $75,0000. Here the differences are evident, with many more tracts in Cluster 3 having high income compared with the other clusters. Cluster 4 also appears to have a relatively small number of high-income tracts. The counterpart of this is shown in Figure 13 for the count below poverty level, where most of the clusters have a relatively flat distribution while Cluster 3 is more concentrated near zero.

As to owner-occupied housing (Figure 14), Clusters 0 and 1 (characteristic of Central and West Central

Virginia) show a lower prevalence of owner-occupied housing compared with Clusters 2, 3, and 4. It may be that the data are showing a balance of tenure (mobility rates differ across regions, as reflected in the demand for more permanent housing) and income (wealthier households are more likely to own their homes).

Turning to data for insurance coverage (Figure 15), Cluster 2 shows a substantial right-skew showing a substantial number of households that work full time but do not have health insurance coverage. That said, there are a surprising number in each cluster lacking health insurance. Figure 17 shows data for Medicaid coverage (targeted to low-income persons), with all the clusters having a number of persons under Medicaid, albeit with more right-skew for clusters 1 and 3.

Lastly turning to computer usage and coverage, Figure 16 shows the counts for broadband of any type across clusters. Broadband is clearly more prevalent in Clusters 0 and 1, although it is notable that even in the other clusters, only a modest share of households have no broadband at all. There are only two tracts in Cluster 2, 3 in Cluster 3 and two in Cluster 4 that report no broadband.

Conclusion

I identify 5 distinct socio-economic clusters in Virginia, using data from the US Census Bureau and the Yelp API. The clusters have distinct distributions for almost all the variables of interest. Each cluster is more prevalent in some regions of Virginia but is present in others as well. This shows that the regions of Virginia are not as homogeneous as might be thought at first–the clusters that predominate in Northern Virginia exist in Southside and Southwest Virginia as well, and vice versa. That is, while Virginians fall into distinct socioeconomic categories, these categories are not uniquely region-specific. This mirrors the finding in other studies that show socio-political diversity even at a granular geographic level. it also contradicts the narrative that characterizes broad regions with an equally broad brush.

The findings have a number of potential applications–including marketing, political organization, social enterprises, and provision of public services. Specifics of these applications await further study.

References

[Caren,] Caren, N. Sushi Bars and Buffets: Measuring local culture with the Yelp API (nealcaren.github.io).

[Census,] Census, B. Census tracts (https://www2.census.gov/geo/pdfs/education/censustracts.pdf).

[Gabriel, 2023] Gabriel, T. (2023). 'Virginia is the Test Case': Youngkin Pushes for G.O.P. Takeover This Fall. *New York Times*.

[McCarty et al., 2018] McCarty, N., Rodden, J., Shoir, B., Tausanovitch, C., and Warshaw, C. (2018). Geography, uncertainty and polarization. *Political Science Research and Methods*.

[Myrick, 2023] Myrick, S. (September 18, 2023). Virginia's elections are a bellwether for conservative policy nationwide. *The Hill*.

[Rodden, 2019] Rodden, J. (2019). *Why Cities Lose*. Basic Books.

[Scholz and Stephens, 1987] Scholz, F. and Stephens, M. (1987). K-Sample Anderson-Darling Tests. *Journal of the American Statistical Association*, 82(399).

[Thaler, 1994] Thaler, R. (1994). *The Winner's Curse: Paradoxes and Anomalies of Economic Life*. Princeton University Press.

[Wasserman, 2014] Wasserman, D. (October 8, 2014). Senate Control Could Come Down To Whole Foods vs. Cracker Barrel. *FiveThirtyEight*.

[WeldonCooper and UVa.,] WeldonCooper and UVa. Virginia's demographic regions (https://demographics.coopercenter.org/virginia-regions).

[Wezerek et al., 2021] Wezerek, G., Enos, R. D., and Brown, J. (May 3, 2021). Do You Live in a Political Bubble? *New York Times*.

[Yancey, 2023] Yancey, D. (October 11, 2023). 10 important things about this year's Virginia elections. *Cardinal News*.

[Yelp,] Yelp. Getting Started with the Yelp Fusion API (https://docs.developer.yelp.com/docs/fusion-intro).

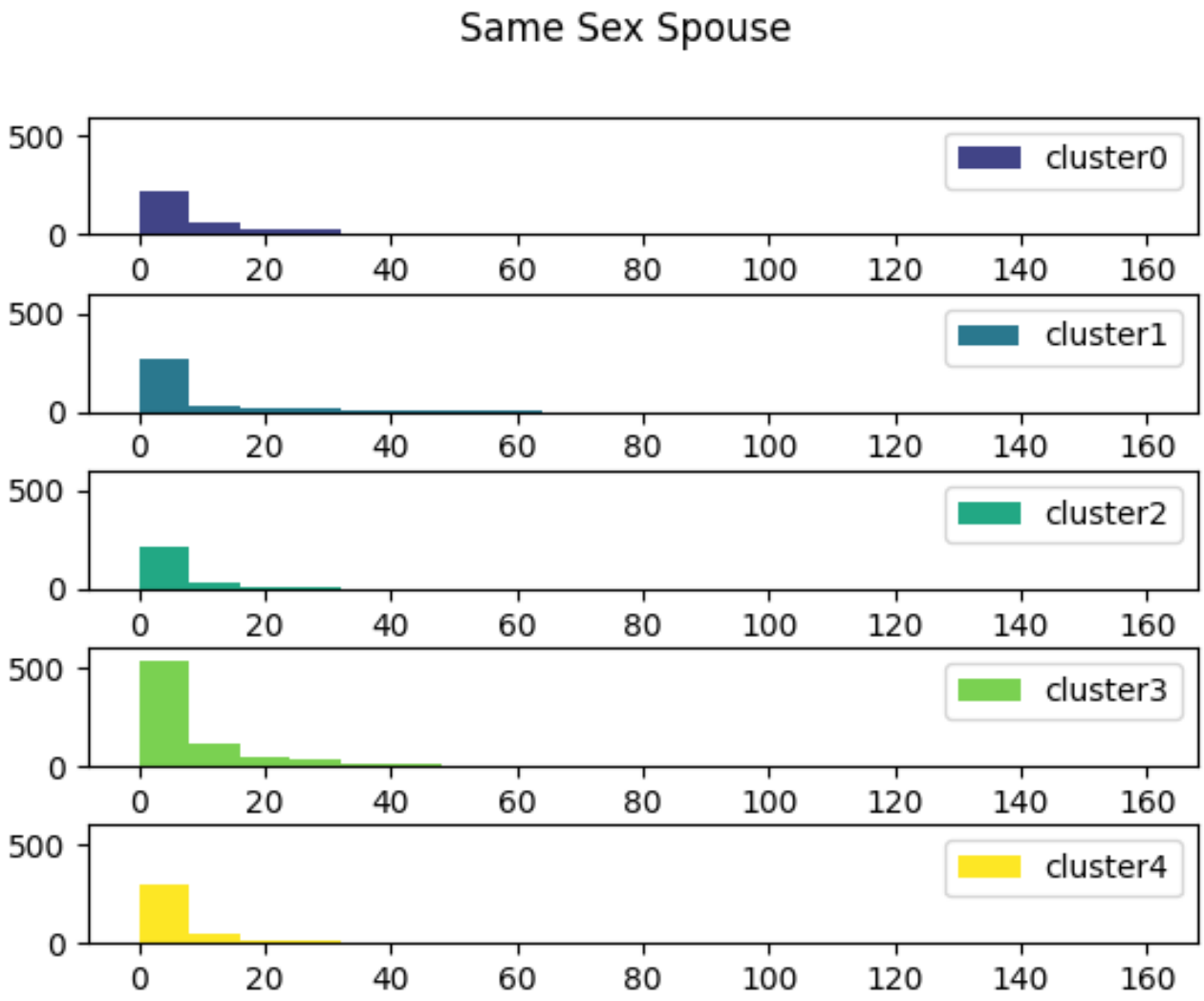


Figure 11: Distribution Across Clusters

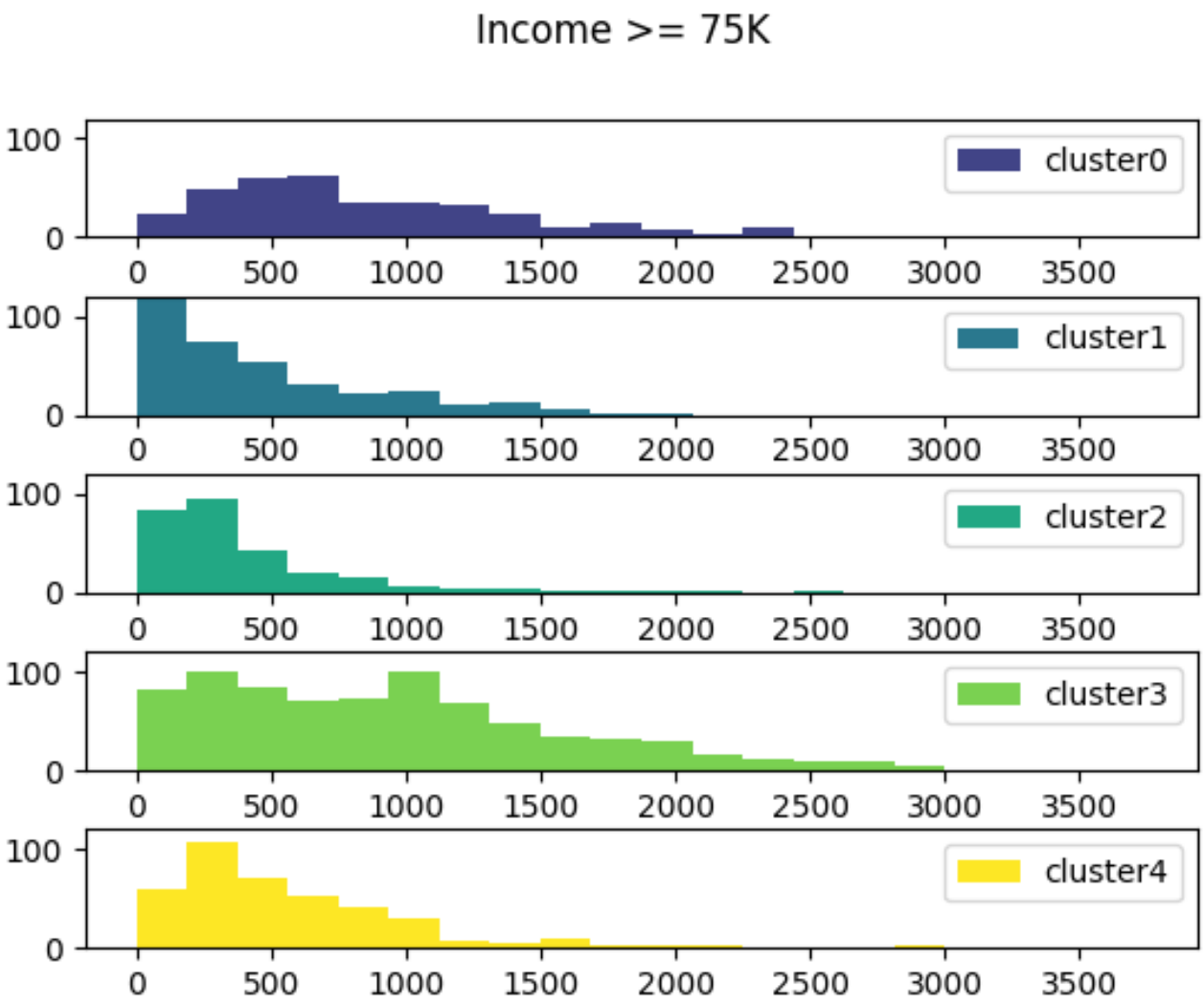


Figure 12: Distribution Across Clusters

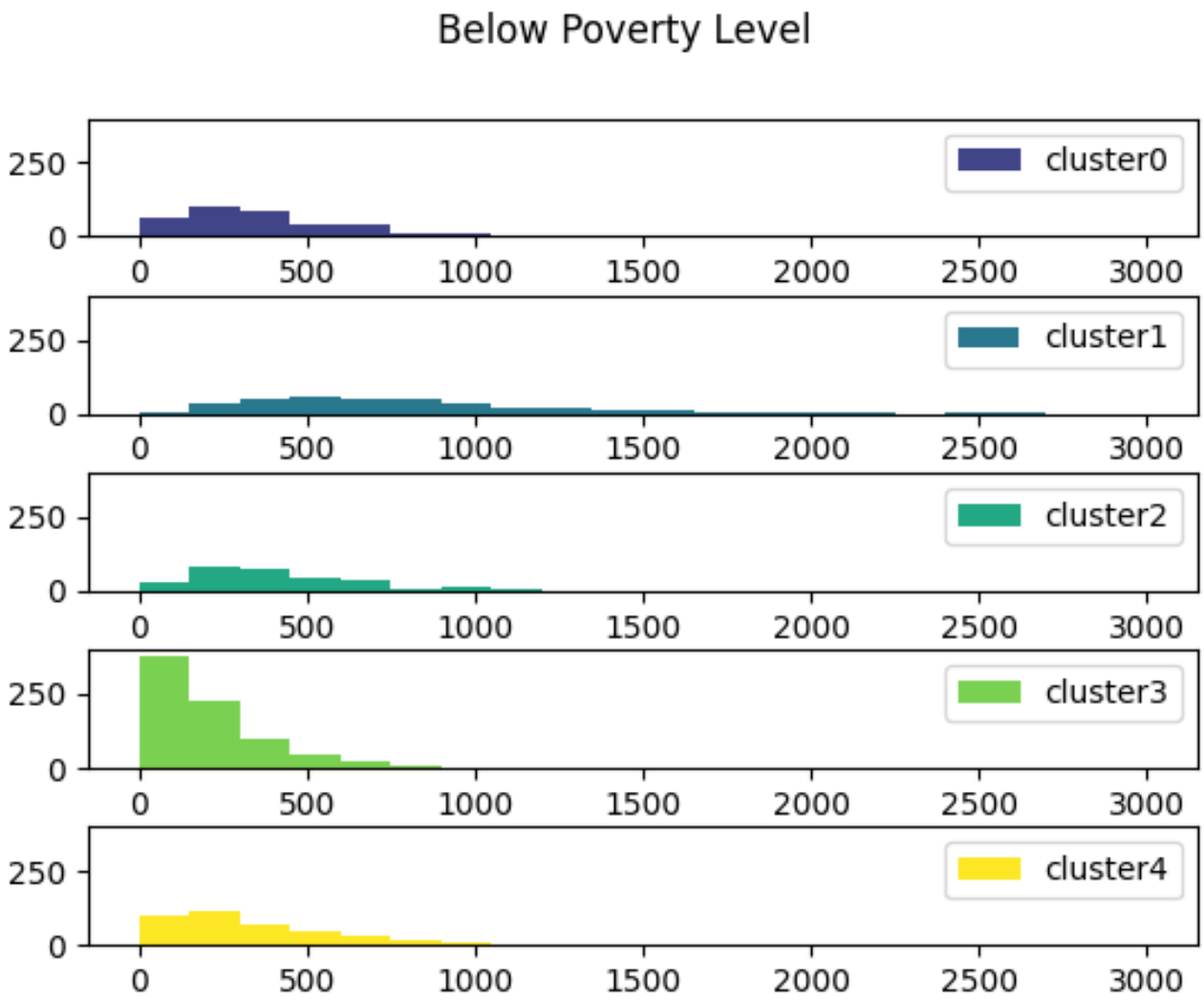


Figure 13: Distribution Across Clusters

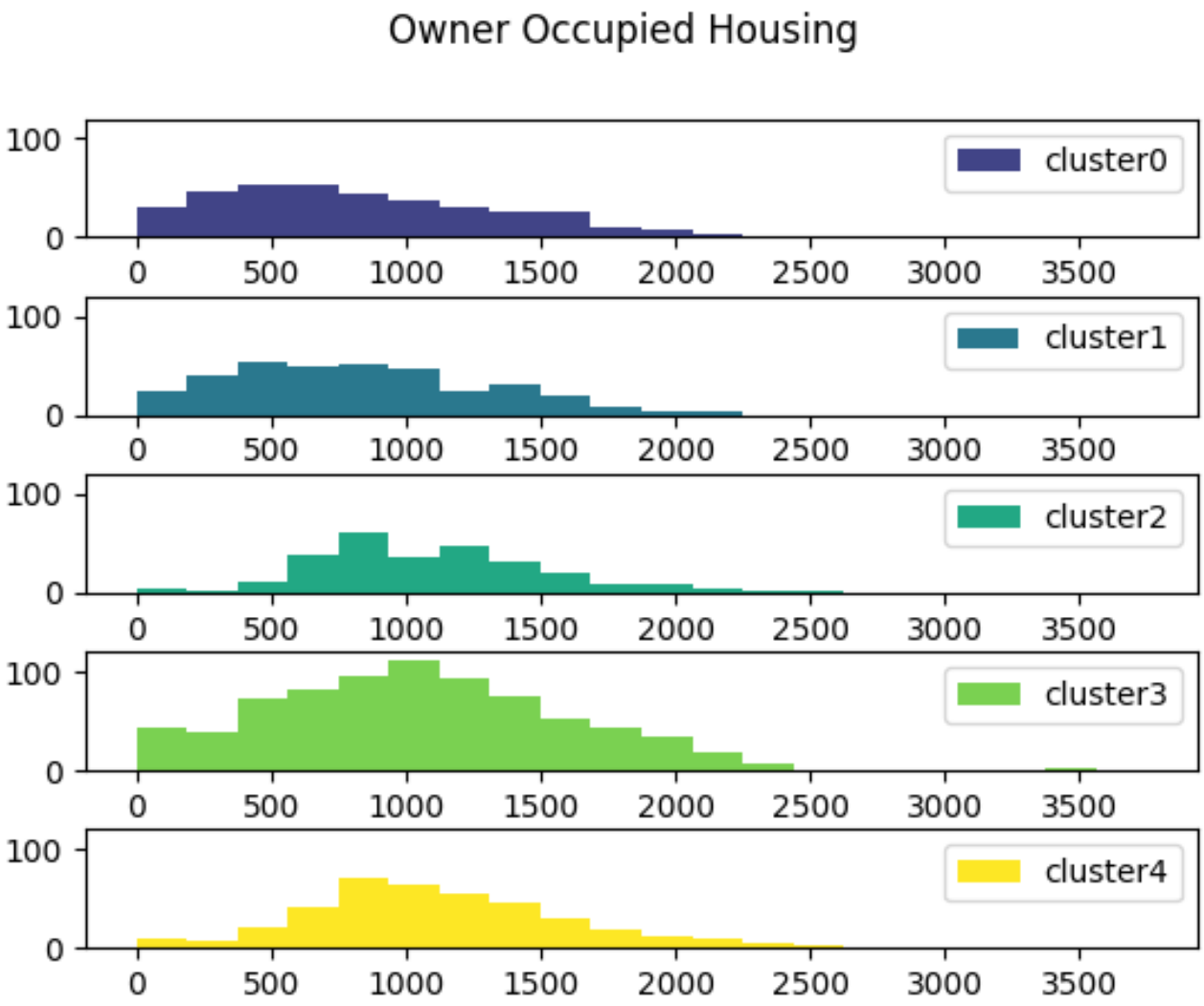


Figure 14: Distribution Across Clusters

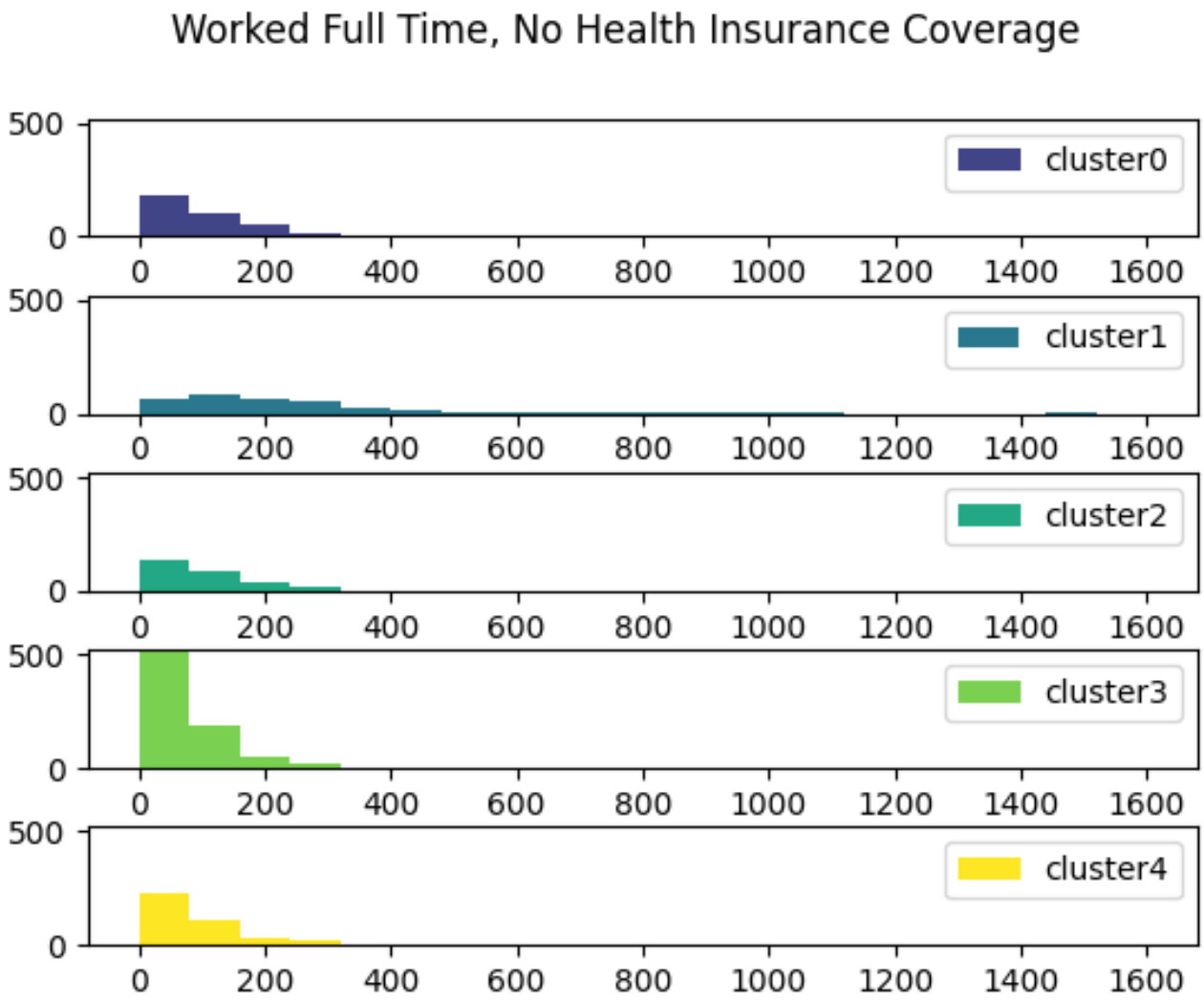


Figure 15: Distribution Across Clusters

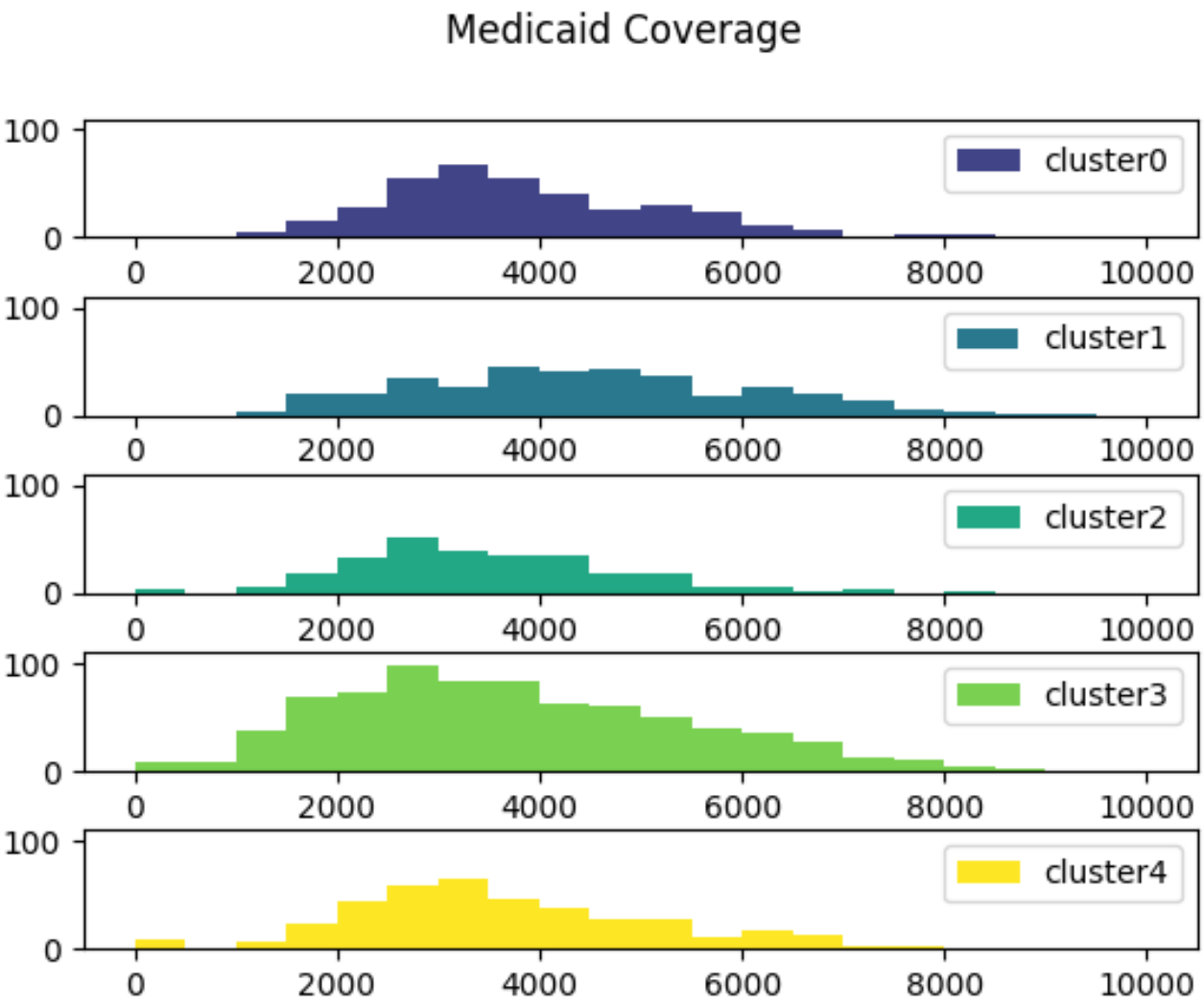


Figure 16: Distribution Across Clusters

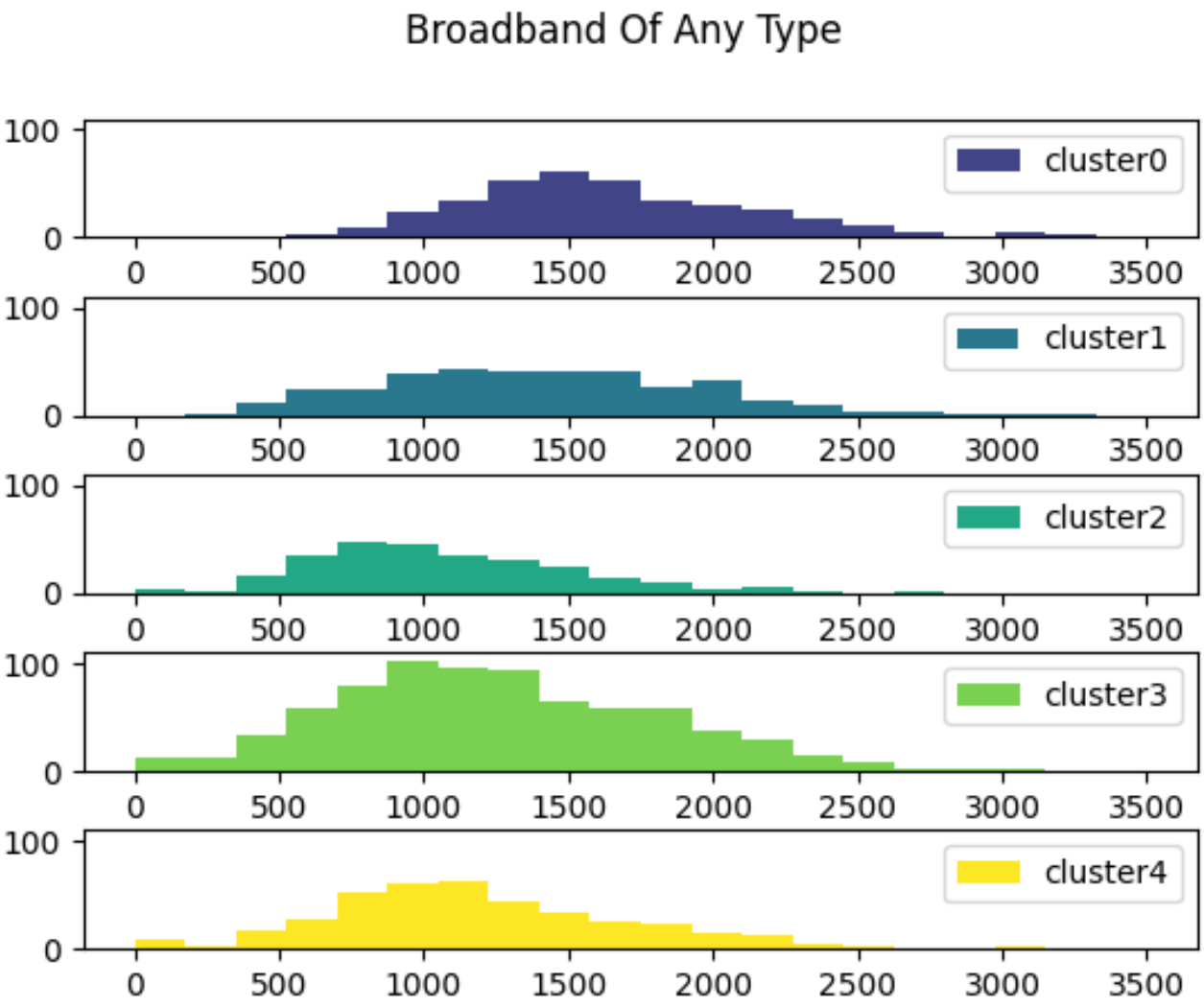


Figure 17: Distribution Across Clusters