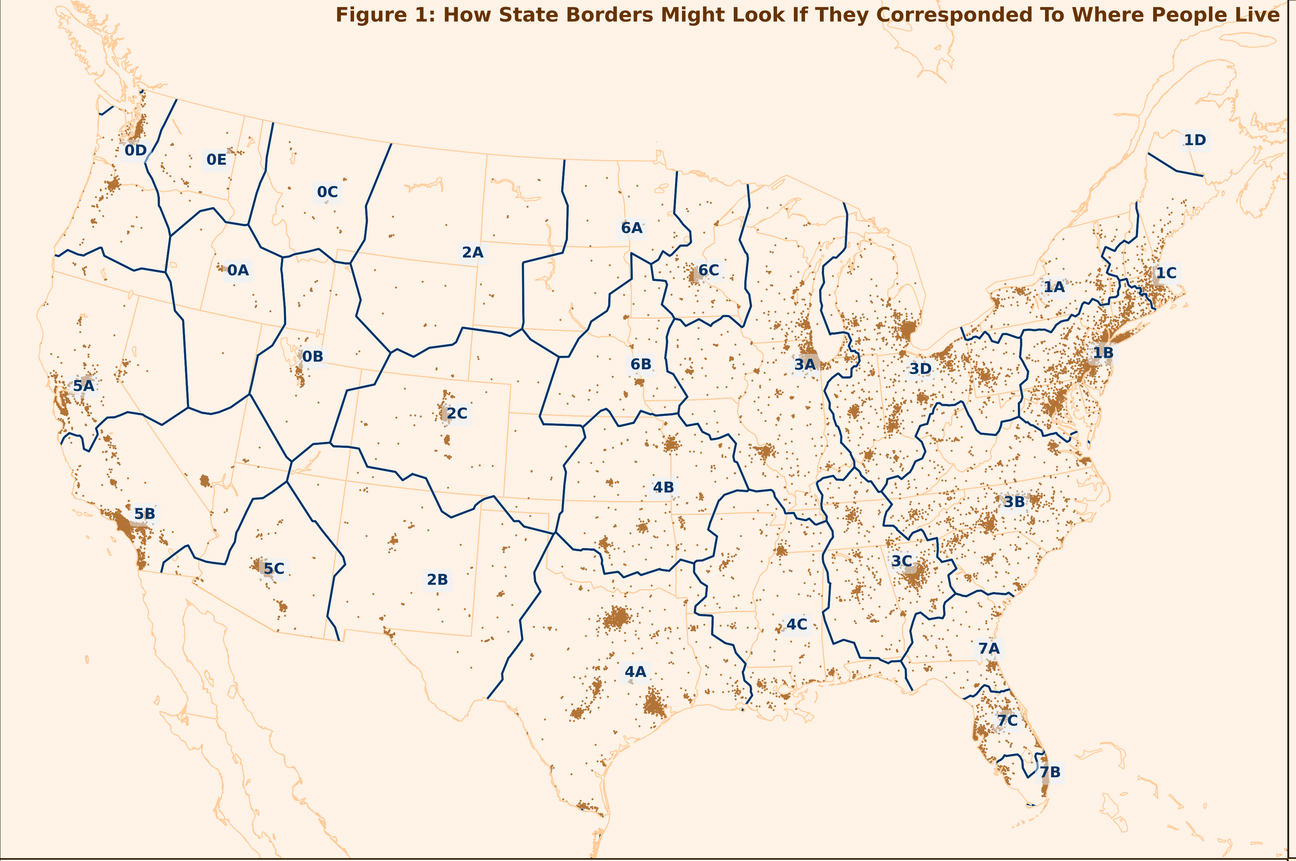
Figure 1: How State Borders Might Look If They Corresponded To Where People Live



States are the preeminent geopolitical units of the United States. Americans live within different public institutions and laws in each state. However, state boundaries are counterintuitive, given their importance. For example, in Kansas City, the Missouri-Kansas boundary cuts the metropolitan area in half. In Texas, El Paso is 285 miles from the nearest metropolitan area in Texas (Odessa) but only 45 miles from the closest one in New Mexico (Las Cruces).

What if the state boundaries of the United States matched where Americans live? This project applies machine learning to reimagine the borders of the contiguous United States. The algorithm groups people who live near each other, placing the new state borders in unpopulated expanses of land between groups. The algorithm suggests grouping the contiguous US population into 28 new states. Figure 1 draws boundaries (blue lines) around population centers (brown dots) to illustrate how the new state borders correspond to where people live.

How would the new states differ in typical quality of life and politics? Figures 2a, 3a, and 4a present the Human Development Index (HDI), total population, and percentage that voted for the 2020 Republican presidential candidate for each of the new states. HDI is a quality of life score that considers high income per capita, long average lifespan, and high average years of education to be signs of high quality of life. The most recent presidential vote is an indicator of general political inclinations. The new states are ordered the same way in all three bar charts to facilitate side-by-side comparisons. Figures 2b, 3b, and 4b present these statistics as a map.

Figure 2: The median new state HDI score is 815, and all new state HDI scores would be considered 'high' (>700) or 'very high' (>800) by global standards. The new states containing Seattle WA, San Francisco CA, Denver CO, Minneapolis MN, and the urban corridor from Boston MA to Washington DC ("Bos-Wash") have particularly high scores.

Figure 3: The median new state has a total population of 6.8 million. However, new states vary widely in population. The top 5 new states have a combined total population of 148.1 million, outnumbering the other 23 states. There is not a strong correspondence between population size and HDI – some populous new states have high HDI scores, while others do not.

Figure 4: In the median new state, 50 percent of voters chose the 2020 Republican candidate. Of the 28 new states, 8 would lean towards republicans, and 10 would lean towards democrats. HDI corresponds strongly with republican lean. The median HDI among republican-leaning new states is 813, compared to a median of 832 among democrat-leaning new states.

Fig. 2a: HDI Score For Each Cluster

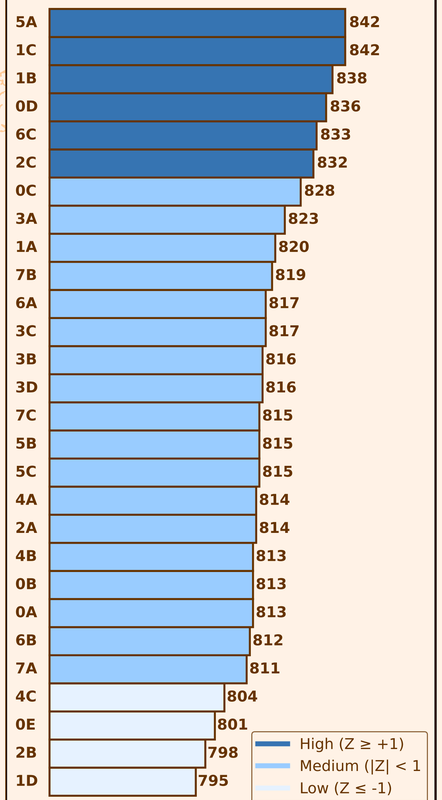


Fig. 3a: Population Total For Each Cluster

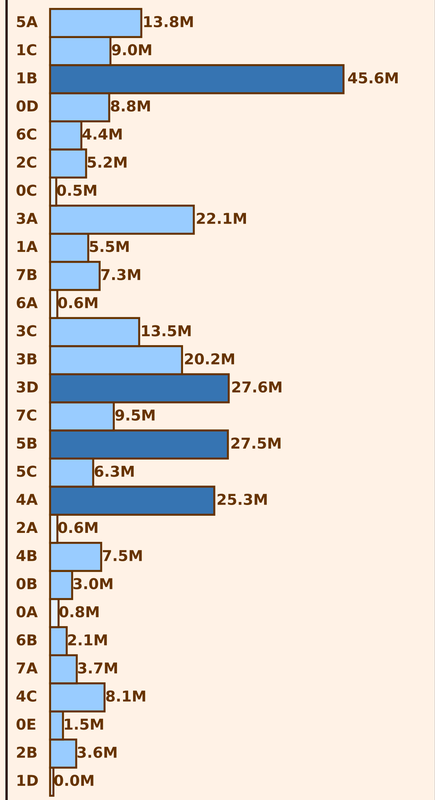


Fig. 4a: Percent Voting For The Republican Pres. Candidate in 2020

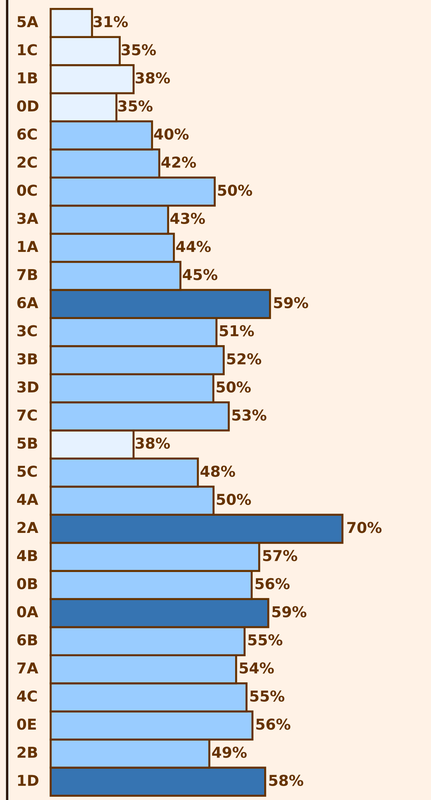


Fig. 2b: HDI Quantile Map

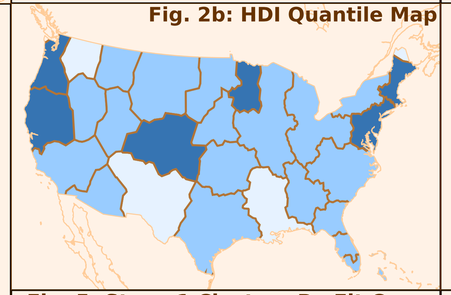


Fig. 3b: Population Quantile Map

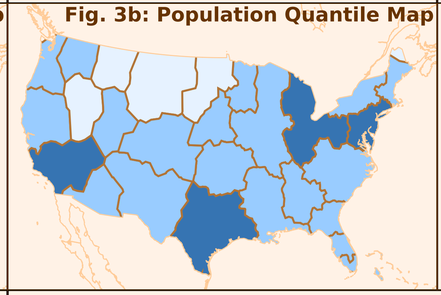


Fig. 4b: Republican Quantile Map

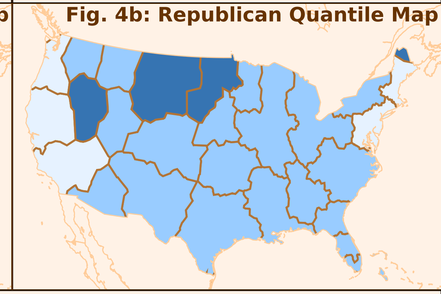
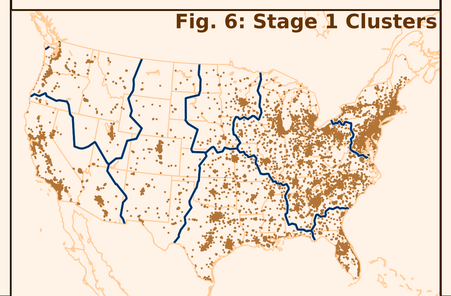


Fig. 5: Stage 1 Clusters-By-Fit Curve

[TO BE ADDED]

Fig. 6: Stage 1 Clusters



Method: Calculating New State Borders

The project applies two-stage agglomerative clustering to formulate new state borders, selecting the number of clusters with a clusters-by-fit curve. Agglomerative clustering is a machine learning technique that groups places into clusters. The algorithm initially groups neighboring places and then iteratively adds more sites to each cluster, starting with sites that are, on average, very close to all the places already in the cluster. It also merges nearby clusters until reaching a specified number of clusters.

Clusters-by-fit curve: Clusters are high-quality if the average distance between places in the cluster is low. Plotting the number of clusters against the quality of those clusters ("fit") provides a principled way to select how many clusters are ideal. One way to determine the best solution is to use this curve to find the number of clusters that provides the highest quality for the least number of clusters – the simplest, high-quality solution. Figure 5 shows the clusters-by-fit curve for the first stage, and Figure 6 shows the clusters selected.

Two-stage process: A single round of clustering produces new state borders dramatically larger than even the largest contiguous US state (See Fig. 6). Project repeats the same clustering process to divide each Stage 1 cluster into smaller Stage 2 clusters.