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.