Bridging the Gap between Citizens and Local Governments: A Democratized Data Solution

Erick Garcia, Saisuriya Paranthaman, Viet Nguyen, Yuchen Wang, Binh Vu, Yong Yan Zhu
Georgia Institute of Technology

Summary

How does the government determine where aid should be distributed? Current solutions include quantitative easing and identifying individuals/families of need based on unemployment insurance and Earned Income Tax Credit (EITC), etc. However, all these solutions have its drawbacks and typically aid distribution is not transparent. here we propose a new solution. We've used a clustering algorithm that will utilize income tax statistics data to identify regions that pay the most taxes and developed a democratized tool with the readout to provide equal access to data for both local governments and citizens.

Data

Our two datasets are on the zip code level:

a) IRS- 2020 Individual Income Tax Statistics, **around 16,500 observations**
b) USAspending.gov – 2020 Financial Assistance, **more than 1,000,000 data points.**

To compare zip different zip code incomes, we calculated the weighted income average for each zip code. Additionally, we used the same equation to calculate four other predictors (unemployment, investment, state tax, and real estate tax). Totaling 5 distinct predictors.

$$\sum_{b=1}^{b}(\text{Income Amount}_b * \text{Num of return}_b) / \sum_{b=1}^{b}(\text{Num of return}_b)$$

K-means algorithm

K-means is an unsupervised clustering algorithm that we used for **unbiased segmentation** of income data. K-means iteratively assign cluster classification to each zip code observation based on the Euclidean distance to the closest cluster.

$$CP(x_1, x_2, ..., x_k) = (\frac{\sum_k^{i=1} x^{1st}}{k}, \frac{\sum_k^{i=1} x^{2nd}}{k}, ..., \frac{\sum_k^{i=1} x^{nth}}{k})$$

With the cleaned income dataset, we ran a wide range of cluster numbers with every iteration adding more and more income data attributes. In the end, we found all attributes to be significant and use it to train the final k-means model.

Visualization

Our tool in Tableau allows users to explore clustered data (**2-20 clusters)**, apply human intuition, and identify areas that require aid, potentially leading to more equal policies.

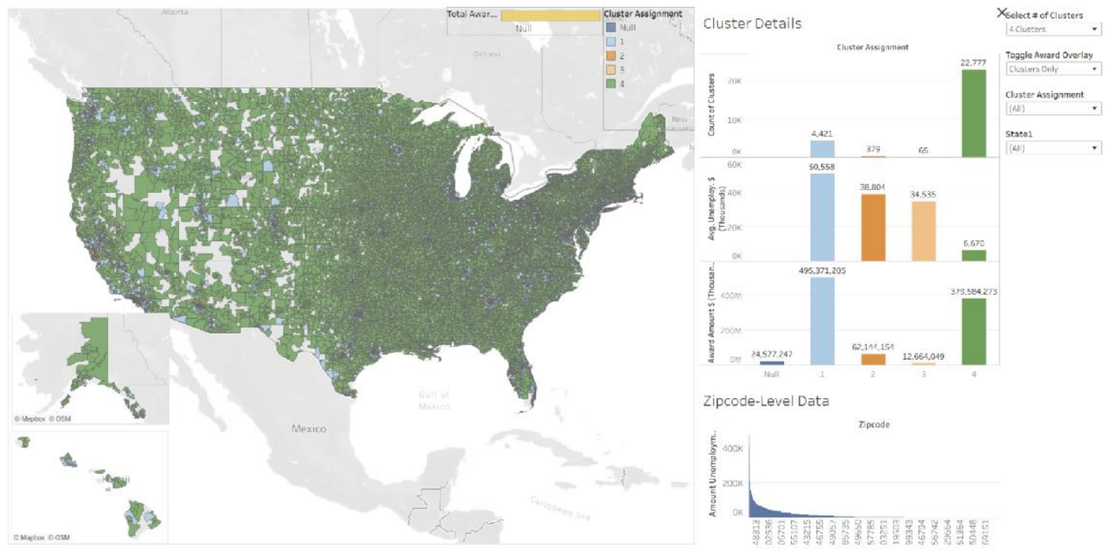

Exhibit #2: "Clusters Only" view with Select # of Clusters (4)

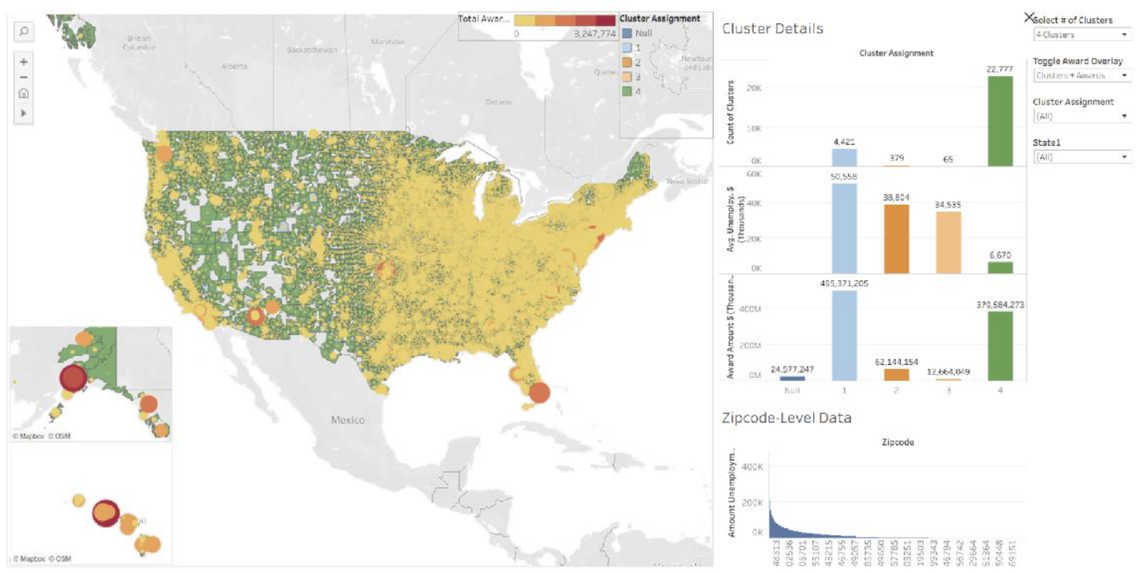

Exhibit #3: "Clusters + Awards" view with Select # Clusters (4)

Experiments and Results

Our tool **increases the transparency of aid distributions** and can have potential benefits such as increasing community participation and enhancing tax administration and compliance. To evaluate our approach, we tested different cluster inputs and reviewed the visualization output until we observed potentially meaningful clusters. Compared to other methods, we intentionally did not assign classifications that told the user which areas did/did not require aid. This allows the user to independently explore patterns in the tax and award data to find new areas that require financial support.

We acknowledge drawbacks such as the need for additional data sources and potential bias in the clustering process. Further studies are advised to address these drawbacks and expand the tool's applicability to different fields and funding sources.