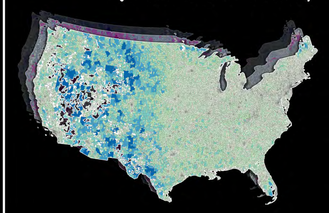
# The U.S. in 11 Zip Codes: Mapping Socio-Economic Clusters & Exemplars of the United States

Nicholas Bogen, Dept. of Geography and Environmental Studies and Center for Geographic Information Sciences
Central Michigan University, Mt. Pleasant, Michigan 48859

CMU Central Michigan University | College of Science & Engineering

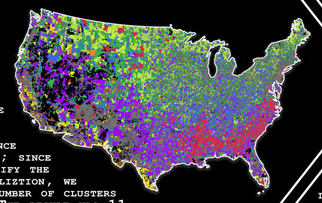
## 1. Data Collection & Preprocesing (ArcGIS Pro, Excel, R)



We acquired data from the US Decadal Census (2010) and American Community Survey (2008-2012) at the Zip Code Tabulation Areas (ZCTAs) for 40 variables that described population, education, income, ethnicity and race, language, and employment. There were a total of 32,115 ZTCAs after removing ZTCAs that had missing data for any of the 40 variables. To reduce the dimensionality of this data and standardize values, principle components analysis (PCA) was conducted in R; 26 components were used representing over 95% of the original data variance.

## 2. Affinity Propagation (R)

Affinity Propagation (AP) is a statistical clustering algorithm that considers each data point as a potential cluster exemplar (Frey & Dueck, 2007). The ZTCA data from step 1 was convirted into a dissimilarity matrix using the weighted Euclidean distance between each ZTCA. AP has a single parameter, the preference value, which affects the relative number of output clusters. Using the median preference value resulted in over 200 clusters; since the goal of this project was to simplify the dimensionality of the data for geovisualiztion, we reduced the preference value until the number of clusters fell into a desired range (10 – 14). The result was 11 clusters and the exemplar ZTCA for each cluster.



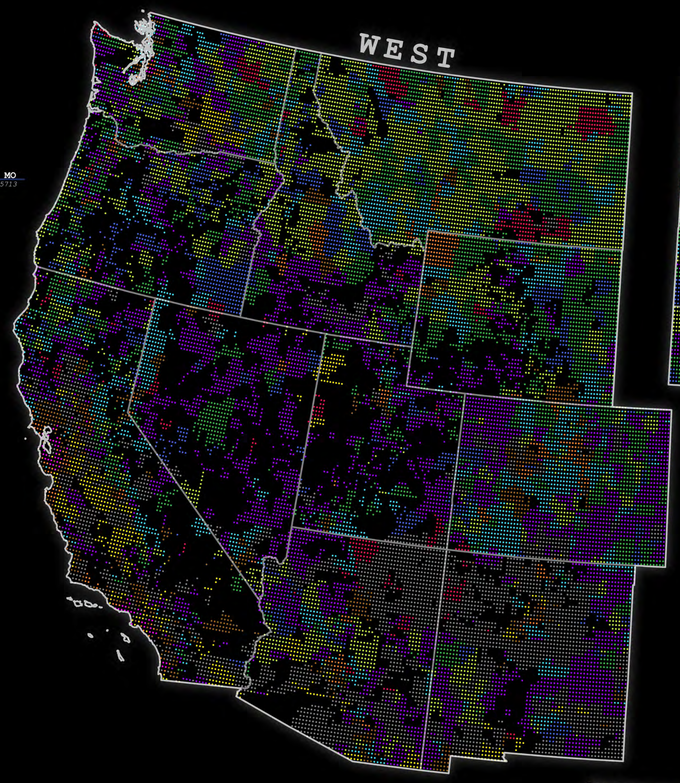
## 3. Tessellation of Exemplars (ArcGIS Pro)

Our clusters and exemplars were mapped to ZCTA; However, mapping ZCTA presents many challenges including a large variation in the size of each ZCTA which over emphasizes large rural western ZCTA and hides small urban ZTCAs. In order toaddress this issue, a square tessellation (100 sq. mi) was created and a majority filter was used to generalize the variably sized ZCTA data into an equal area grid. For aesthetic purposes, the square tessellation was converted in to a point file based on the centroid of each square. These points were then mapped using a 11-color qualitative color scheme on a black background to aid in the visualization of a large number of non-hierarchical classes.

## 4. Point Density Mapping (ArcGIS Pro)

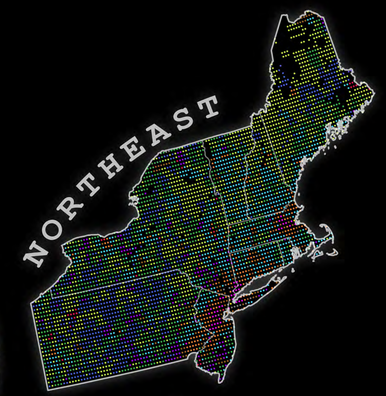
To further enhance the visualization and interpretation of the data, each individual cluster and its exemplar were mapped. A point density map was used to highlight areas where a given clusters are predominant and a callout for each exemplar was created listing the ZCTA and administrative name.

**Blue** indicates:
- Avg. Age Distribution
- Higher % of pop. w/ only High School edu.
- Slightly higher than avg. White pop.
- Slightly lower than avg. Black pop.
- Highest Native born pop.
- Highest English speaking only pop.









**Pink** indicates:
- Lowest % of pop. 20-24 Y.O.
- Higher than avg. % of pop. 25-64 Y.O.
- Extremely wealthy exemplar
- Highest Per Capita Income
- Highest % pop. with incomes of $200,000.
- Sig. higher than avg. jobs in management, business, science, and arts jobs.
- Sig. lower number of jobs in other job categories.
- Extremely higher than avg. % of pop. w/ Bachelor's or higher.

**Lime Green** indicates:
- Lowest % of 19 Y.O. or younger pop.
- Highest % of 65 Y.O. or older pop.
- Highest White & Lowest Black pop.
- High % of Foreign born from Europe or North America.
- Slightly higher than avg. Native born pop.

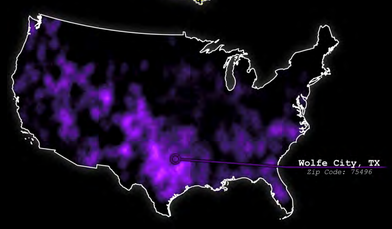
**Yellow** indicates:
- Lower than avg. white pop.
- Higher than avg. pop. of 19 Y.O. or younger
- Extremely higher than avg. Hispanic/Lantino pop.
- Sig. lower than avg. 65 Y.O. or older pop.
- Extremely higher than avg. family size.
- Higher than avg. % of pop. below poverty level.
- Extremely higher than avg. Spanish speaking pop.

**Orange** indicates:
- Higher than avg. pop. of 20-24 Y.O.
- Sig. higher than avg. Per Capital Income
- Sig. higher than avg. % of pop. with income $200,000+
- High # of pop. with jobs in management, business, science, and arts.
- Extremely higher than avg. % of pop. w/ Bachelors or higher.
- Fairly avg. race identity distribution.

**Gray** indicates:
- High % of pop. 19 Y.O. or younger.
- Higher than avg. Hispanic/Latino pop.
- Higher than avg. family size.
- Highest pop. of service, sales, and office jobs.
- Lower than avg. married pop.
- Lower than avg. White pop.
- Higher than avg. Spanish speaking pop.

**Green** indicates:
- Lowest % of pop. with incomes of $200,000+.
- Avg. Income Per Capita.
- Slightly higher than avg. White pop.
- Lowest % pop. of Nat. Amer.

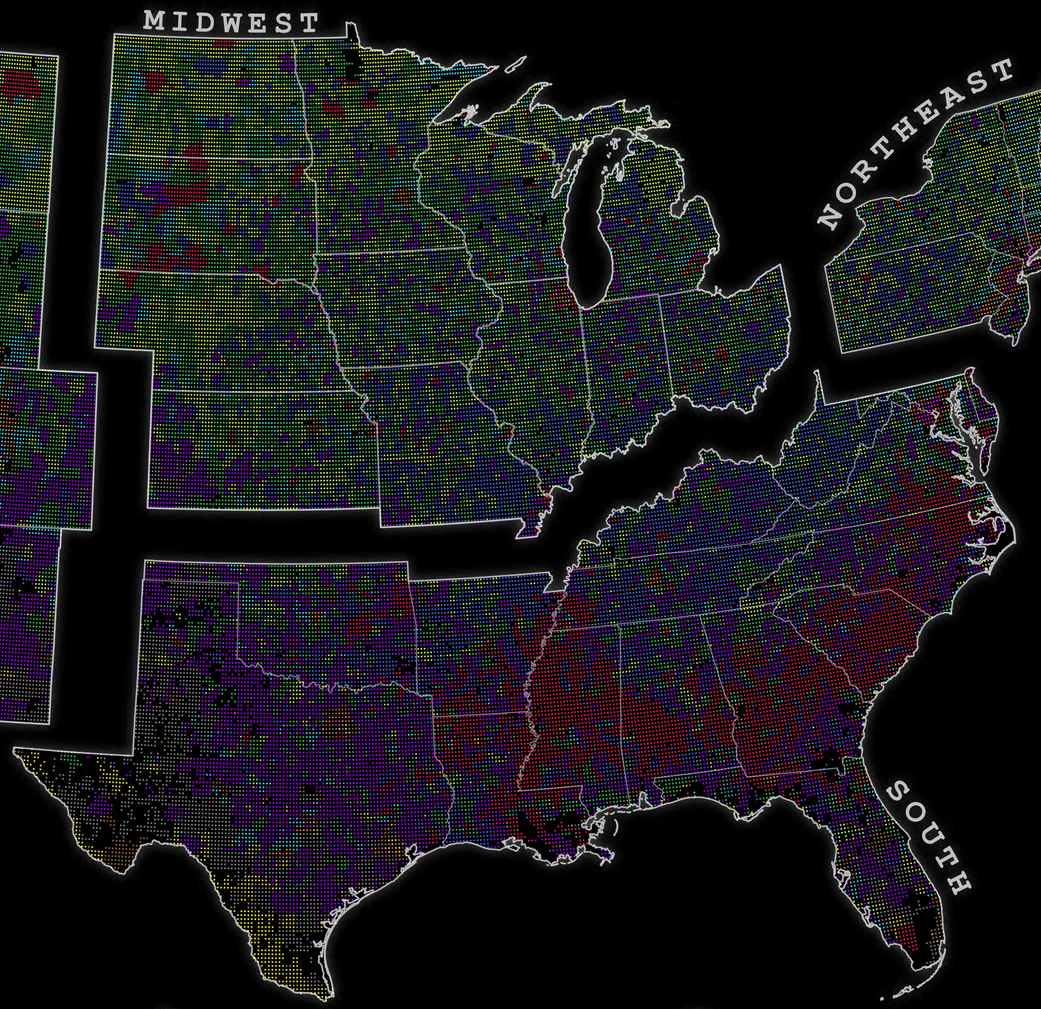
**Cyan** indicates:
- Slightly higher than avg. Per Capita Income.
- Highest married pop.
- Slightly higher than avg. % of pop. w/ Bachelor's or higher.




**Purple** indicates the most average exemplars (very few categories of the population make-up are above or below .05 Standard Deviations from the mean) and has a slightly higher avg. pop. of foreign born Latin Americans.

**Brown** indicates:
- Slightly higher than avg. pop. of 20-24 & 25-64 Y.O.
- Slightly higher than avg. pop w/ Bachelor's or higher.
- Slightly higher than avg. Per Capita Income.
- Above avg. pop. that speak other Indo-European Languages.

**Red** indicates:
- Lowest married pop.
- Sig. lower than avg. White pop.
- Sig. higher than avg. Black and Native American pop.
- Lower than avg. % of pop. below the poverty line.




Contact Information: Nicholas Bogen • bogen1nk@cmich.edu | Data Source: US Census Bureau and The Central Michigan University Center for Geographic Information Science | Acknowledgements: Partial funding provided by Central Michigan University College of Science & Engineering and the Central Michigan University Office of Research & Graduate Studies. Data collection, Running of Affinity Propagation, and Interpretation of Data conducted by Benjamin Huemann, Matthew Liesch, David Patton, and Ryan Meier. Cited Affinity Propagation Research Publication: Frey, B. J., and Dueck, D. (2007) Clustering by passing messages between data points. Science 315, 972-976 | Note that the above maps are not to scale.