The contiguous United States in eleven zip codes: identifying and mapping socio-economic census data clusters and exemplars using affinity propagation

Benjamin W. Heumann\textsuperscript{a,b}, Matthew E. Liesch\textsuperscript{a}, Nicholas R. Bogen\textsuperscript{a,b}, Ryan A. Meier\textsuperscript{a,b} and Marcello Graziano\textsuperscript{a,b}

\textsuperscript{a}Dept. of Geography and Environmental Studies, Central Michigan University, Mt. Pleasant, MI, USA; \textsuperscript{b}Center for Geographic Information Science, Central Michigan University, Mt. Pleasant, MI, USA

ABSTRACT
The United States is a diverse and heterogeneous place. Accurately organizing and mapping the U.S. into different regions based on characteristics such as wealth, race, education, language, and occupation is a complicated and arduous task. This paper demonstrates the application of affinity propagation to map socio-economic patterns and identify representative exemplars. Affinity propagation clusters data based on representative exemplars and considers all data points as potential cluster exemplars. We use socio-economic data from the United States census to cluster zip codes tabulation areas and identify representative locations of socio-economic diversity of the United States. The 11 socio-economic clusters were mapped individually and together using area-based generalization. Mapping the results illustrated distinct regionalization and historical migration trends within the United States as well as national urban/suburban/rural patterns. Future applications of this technique may be useful for data-driven socio-economic analysis and purposive sampling.

1. Introduction
The identification, investigation, and comparison of ‘regions’ is one of the most recurring themes throughout geography (Hazzi et al., 2018; Paasi 2002; Sidaway & Hall, 2018; Weaver & Holtkamp, 2016). Countries are normally diverse and heterogeneous agglomerations of places, whose differences and similarities span across various metrics (Culcasi, 2010; Liesch, 2008; Semian, 2016). Accurately organizing countries into different regions based on characteristics of wealth, race, education, language, and occupation is a complicated and arduous task (Culcasi, 2010; Sidaway & Hall, 2018). Other scholars have used these and similar demographic categories to create typologies for different socioeconomic regionalizations of rurality in Europe (Dicka et al., 2019; Hedlund, 2016; Van Eupen et al., 2012). Socioeconomic data illuminates patterns previously harder to spot; regions derived this way normally have different spatial extents than what vernacular regional boundaries might suggest (Chinni & Gimpel, 2010; Hedlund, 2016). In so doing, socioeconomically-delineated groupings of places make possible new research questions and further, can inform policy application with regard to specific socioeconomic areas (Casellas & Galley, 1999; Dicka et al., 2019; Van Eupen et al., 2012). The organization of socio-economic data centers around two objectives: grouping similar administrative units together and the selection of representative administrative units for each group. Knowing the locations that best represent the U.S. and its people can be beneficial to scholars, policymakers, and market analysts alike (Casellas & Galley, 1999; Green et al., 1967), especially for mixed methods research in which only a few typical or representative locations are examined (Teddlie and Yu, 2007).

The goal of clustering or selecting representative communities can be traced back to Lynd and Lynd (1929), whose classic study sought to explore lifestyles within a typical American community. In their study, the authors used basic demographic and observation data to identify Muncie, IN as ‘Middletown’, a city that shared characteristics with the widest group of communities.

Clustering analysis, a form of data mining, involves the analysis and extraction of patterns previously unknown due to the size of the data set (Kopanakis & Theodoulidis, 2003). Jonathan Robbin was one of the first to employ geodemographics, a subfield which uses certain social, economic, and behavioral statistics to create a prediction model for marketers to identify geographic areas that are best suited for their product (Goss, 1995; Robbin, 1980). He created a computer program (PRIZM) that used census data and consumer surveys to group U.S. zip codes into forty lifestyle clusters (Weiss, 1989). Weiss (1989)
followed up on Robinson’s work by visiting a place in each of the forty clusters to interview members of the community and determine of the accuracy of Robin’s program. Weiss (1989) concluded that the majority of locations he visited matched the cluster description of which they were a part. Similar contemporary research has focused on identifying and mapping types or classes of communities (Brookfield et al., 2005), cities (e.g. Bibby & Brindley, 2017; Bolton & Hildret, 2013), counties Chinni and Gimpel (2010) and rural areas (Copus, 2015), whether using multi-criteria, user-defined indexes, pattern-seeking techniques, or user-defined threshold-base characterizations.

Cardille and Lambois (2010) used cluster analysis to identify signature landscapes of the continental U.S. Their purpose was to aid an increasing interest of ecosystem management and other research that would benefit from finding study areas that exemplify land-use/land-cover (LU/LC) distribution at larger sub-national regional scales. The algorithm resulted in 17 distinct landscapes, recognizing the role of human influence, which can be considered as the most expressive landscapes of the continental U.S.A. This cluster analysis was possible due to the application of a novel clustering algorithm, Affinity Propagation (AP).

AP was created in Frey Labs at the University of Toronto as a tool to help identify unforeseen patterns in large data sets by creating subsets of similar features within the data (Dueck & Frey, 2007). It is an effective clustering algorithm for large data sets and outperforms the commonly used k-centers algorithm (Frey & Dueck, 2007). AP considers all data points as possible exemplars and iterates through all scenarios until the sum of the dissimilarity between each data point and the exemplar of their cluster is the least. This process is similar to the Fisher-Jenks algorithm for optimal classification in that every possible grouping or class break is considered until the ideal is determined (Slocum, 2009). A major difference is that AP does not restrict the grouping into a predefined number of clusters as the Fisher-Jenks classification does. The number of output clusters for AP is influenced by a parameter called the preference value. Frey (2009) typically recommends using the median value of the dissimilarities between data points, but can result in a medium to large number of exemplars. For fewer exemplars, Frey (2009) recommends starting with the lowest dissimilarity value, but notes that the preference value can be altered to either increase or decrease the number of clusters. This characteristic gives AP an advantage over other clustering algorithms such as k-centers which groups the data into a predefined number of clusters.

The objective of this paper is to demonstrate the application of affinity propagation to identify, map, and analyze the patterns of socio-economic clusters in the U.S.A. Specifically, we cluster zip codes based on selected United States Census Bureau (hereafter ‘Census’) socio-economic characteristics, identify the exemplar or representative zip codes for each cluster, and examine the distinguishing characteristics of each exemplar. Finally, we illustrate the results using maps of the exemplars and clusters to examine spatial patterns.

2. Methods

2.1. Data

The data used in this project was publicly available 2010 Census and 2008–2012 five-year estimates from their American Community Survey (ACS), partitioned into 32,989 Zip Code Tabulation Areas (ZCTAs) (United States Census Bureau, 2010; United States Census Bureau, 2011). The Census assembled data from 2010 census blocks into 32,989 ZCTAs which are summarized representations of United States Postal Service zip code areas. ZCTAs are based on Census block group boundaries—a collection of census blocks—and do not include large areas with no population or areas of only water. ZCTAs do not have a uniform geographic size. Their geographic extent is based on largely upon population, where larger ones, located in rural areas, contain sparser population, while smaller ZCTAs, primarily located in urban areas, contain denser population. The geographic boundaries of the ZCTAs used in the final map will be provided by the Census website as Topologically Integrated Geographic Encoding Referencing (TIGER) shapefiles.

The selection of enumeration area (i.e. zip codes) was a tradeoff between spatial resolution and data volume. Due to the nature of AP, as the number of data points increase, the amount of space needed for computation increases exponentially. For example, with 32,989 ZCTAs there are 32,989 × 32,989 pairwise dissimilarities. If census tracts were used there would be 73,057 × 73,057 pairwise dissimilarities: about double the number of data points but five times as many dissimilarity pairs. While previous studies have used counties for enumeration units, we chose to avoid that level of aggregations due to their highly variable population and size, especially within urban areas and western states.

Forty attributes were extracted from the decennial census and the ACS that describe the sociodemographic profile of ZCTAs (see Table 1). For this study, we limited attributes to those that described people and population, although additional census and non-census attributes describing the economic activities, land use, or climate could be incorporated. All attributes were continuous values.

2.2. Preprocessing

Approximately 20% of the ZCTAs contained some form of missing or omitted data. Figure 1 illustrates
the decision tree addressing these issues. Some of the missing data was due to uncalculated zero values. For example, if 100% of the population in a zip codes was below the poverty line, the percent income greater than $200,000 was left blank. However, the majority of data omission was due to ‘data for this geographic area cannot be displayed because the number of sample cases is too small.’ (United States Census Bureau / American FactFinder 2010). To ensure consistency within the data, several steps were taken to address these issues. We eliminated 271 ZCTA with zero population. ZCTAs with 100% of the population younger

| Attribute | Source | Attribute | Source |
|-----------|--------|-----------|--------|
| Total Population | 1 | Educational Attainment | 3 |
| Area Square Miles | 2 | Less than High School | 3 |
| % of Population 19 and under | 1 | High School Graduate | 3 |
| % of Population 20–24 | 1 | Some College or Associate’s degree | 3 |
| % of Population 25–64 | 1 | Bachelor’s degree | 3 |
| % of Population 65 and above | 1 | Graduate or Professional degree | 3 |
| % Race Identity & Ethnicity | 1 | Gender | 1 |
| Hispanic or Latino | 1 | % Male Population | 1 |
| White | 1 | % Nativity and Place of Birth | 3 |
| Black or African American | 1 | Native | 3 |
| American Indian and Alaska Native | 1 | Europe | 3 |
| Asian | 1 | Asia | 3 |
| Native Hawaiian and Other Pacific Islander | 1 | Africa | 3 |
| Other Race | 1 | Oceania | 3 |
| Family Size | 1 | Latin America | 3 |
| Average Family Size | 1 | Northern America | 3 |
| Income | 3 | Language Spoken at Home | 3 |
| % Below the Poverty Level | 3 | English Only | 3 |
| Per Capita Income (Dollars) | 3 | Spanish | 3 |
| % Individuals Income => $200,000 | 3 | Other Indo-European Languages | 3 |
| Occupation | 3 | Asian and Pacific Islander Languages | 3 |
| Management, Business, Science, and Arts | 3 | Other Languages | 3 |
| Service | 3 | ""
than 25 with a particular level of educational attainment at zero or too small to report had blank educational values replaced with a 0. Two ZCTAs with 100% of the population younger than 16 had blank values for married and occupation replaced with a 0. 6,174 ZCTAs had 100% native born population with blanks in the foreign born categories and were converted to 0s. 144 ZCTAs with zero population were eliminated from the analysis. 404 (1.22%) ZCTAs were eliminated due to the data omitted due to privacy issues.

Additionally, the study area was restricted to the Continental United States. We acknowledge that both Hawaii and Alaska have unique characteristics not captured by the Census due to their geographic location. Preliminary results that included Alaska and Hawaii found that while Hawaii was largely a distinct cluster, clustering in Alaska was inconsistent often due to large geographic areas inherent in the ZCTAs, that is, ZCTAs were too large because of low population, covering vast, unpopulated territories. In spite of these corrections, 32,115 final ZCTAs existed after preprocessing (97.35%).

In order to standardize the attributes for the dissimilarity calculation, all data were transformed into z-scores. A principal component analysis (PCA) was run on the z-scores using the R statistics software package FactoMineR (Le et al., 2008) to eliminate correlated attributes and reduce data size. The results of PCA found that 94.64% of the data could be represented by just 26 components, a 35% reduction in data size which allowed the entire dataset to be run on a server with 128GB of RAM.

2.3. Affinity propagation (AP)

AP was used to cluster and identify exemplar ZCTAs. The R package APCluster was used to run the actual algorithm (Bodenhofer et al., 2011). AP requires two parameters as input, a dissimilarity matrix and a preference value. The dissimilarity matrix quantitatively defines how different two locations are. The values of the dissimilarity matrix are user-generated and defined. The dissimilarity matrix was generated from the PCA results. A negative weighted Euclidean distance equation based on each component’s eigenvalue percentage of variance was used to calculate the dissimilarity between every point. Equation (1) shows a standard negative weighted Euclidean distance (Greenacre, 2005):

$$\sum_{j=1}^{l} w_j (x_j - y_j)^2$$

where \( l \) is the number of components used from the PCA results and \( w \) is the eigenvalue percentage of variance for each \( j \).

AP cannot accept a predetermined number of clusters as a parameter. Rather, the number of output clusters for AP is influenced by a parameter called the preference value. Typically, the preference value is a common value such as the median of the dissimilarities, although other times it can be the minimum dissimilarity value (Dueck & Frey, 2007). The preference value can be altered to either increase or decrease the number of clusters, but it is not a linear relationship. Initially, the minimum dissimilarity value was selected as the preference value, but it resulted in over 200 clusters. The goal of this work was to produce a more generalized representation of socio-economic zip codes suitable for small sample sizes used in mixed methods; a large number of clusters with few, minor differences from each other, would not serve well for the purposes of generalization and any subsequent detailed research. As a general guideline, the selection of the preference value should match the needs of the specific research, and can be modified to include more or fewer exemplars and clusters. To select the appropriate level, the preference value was doubled until the number of clusters did not change between intervals. The final preference value was −464.51 (22 times the minimum dissimilarity value) resulting in 11 clusters.

2.4. Mapping

Affinity propagation clustering generalized socio-economic census data from 40 continuous variables to 11 clusters and exemplars. The map produced contains a mapped area with all clusters visualized as well as 11 inset maps that illustrate the point-density of each cluster and the location of the exemplar. The colors for the maps were created using Color Brewer (Brewer, 2019) as a starting point, with adjustments suitable to the dark background. Due to the inherent scale issues of ZCTAs, they were generalized by creating a fishnet and assigning the cluster number with the greatest population to a given cell, based on population density of the ZCTA times the area of the given ZCTA in that cell. The gridted data was then converted to centroids for aesthetic purposes. Note that some of the centroids fall outside of the national boundaries as we chose to keep fishnet cells and their centroids as long part of the cell intersected with national landmass. The point density inset maps were generated using the point density tool in ArcGIS Pro using a circle neighborhood with the default radius. The map background is a color desaturated ‘blue marble’ image from NASA. We chose to omit some common map elements such as a north arrow and graticules given how ubiquitous national U.S. maps are in the popular media and these elements would interfere with the aesthetics of the map.
Table 2. List of preference values tested and number of resulting clusters. Bold indicates the preference value selected.

| Preference Value | # of Clusters |
|------------------|---------------|
| 21.11            | 125           |
| 42.23            | 74            |
| 84.46            | 1            |
| 126.69           | 34            |
| 168.91           | 23            |
| 211.14           | 21            |
| 253.37           | 18            |
| 295.6            | 15            |
| 337.83           | 14            |
| 380.06           | 13            |
| 422.29           | 12            |
| **464.51**       | **11**        |
| 506.74           | 11            |

3. Results

3.1. Exemplars

Table 2 lists the 11 exemplars with basic characteristics (population and area) and the standardized z-scores for all of the input variables. While the population of the exemplar ZCTAs ranged from 2,496 (65713 – Niangua, MO, Cluster 5) to 53,697 (92570 – Perris, CA, Cluster 11), the population density was relatively consistent between the exemplars, except Cluster 9 (82009 – Cheyenne, WY). The notable variations have been compiled into a list of distinctive characteristics in Table 4. As Table 3 shows, each exemplar represents several distinct characteristics of the U.S. population based on age distribution, income, wealth, race, language, and occupation. In many cases, socio-economic patterns are correlated and grouped together. For example, 28352 (Laurenburn, NC, Cluster 4), is the exemplar for the cluster that represents the highest levels of poverty, as well as higher percentages of African American and Native American residents. Wolfe, TX, Cluster 7 (75496) is the most average exemplar, with most factors near the statistical mean.

3.2. Maps

The final map illustrating the clusters and exemplars is shown in Figure 2. Many clusters show a strong regionalization. For example, cluster 4, which represents higher levels of poverty and African American and Native American residents, is most prevalent in the ‘Deep South’, with pockets in the American West as well as (post) industrial centers in the Midwest as a result of the early twentieth Century Great Migration from the Deep South to industrial centers in the North (Gibson & Jung, 2002). Clusters 10 and 11, which represent higher percentages of Latin American and Spanish speaking residents, are concentrated in the American Southwest. Cluster 5, which represents areas with higher percentages of white residents with lower education levels, corresponds mostly with the Appalachian and Ozark regions, as well as northern woodlands of the Midwest. Not all clusters have broad regional patterns. Several have spatial patterns that correspond to urban-rural differences in development. Clusters 6 and 9 are found near many urban areas including the Boston–Washington Corridor, the Bay Area, Los Angeles, and major urban centers of the Midwest. Based on the point density maps, we see that the New York metropolitan areas is among the densest areas for several exemplar illustrating both the overall population density of that region, but also highlighting sociodemographic diversity.

4. Discussion and conclusions

The United States is socially and economically diverse, and this diversity manifests itself at varying spatial scales. The generalization of the U.S. helps establish insight into particular characteristics and geographic patterns, while still maintaining an enlightening level of detail. Chinni and Gimpel (2010) presented a portrait of America with 12 community types and revealed many regions and patterns across the U.S. Many regions from this study matched closely with regions in Chinni and Gimpel’s study, including the region across southern states, the region along the Mexico-U.S. border, and the suburbs that stretch between D.C. and Boston. It should be mentioned that this analysis is limited to the data available in the Census; thus it omits some potentially important cultural factors such as religion, politics, social norms, dialect, and voting patterns (Chinni & Gimpel, 2010). Although data does exist for these factors, the finest resolution for this data tends to be at the county level. Counties can be subject to the modifiable areal unit problem because of their inherent inequalities in area and population (Norman, 2006; Norman, 2015; Openshaw, 1984), and that major urban centers normally sprawl across several counties. Additionally, since religion and political data are not part of the Census, their survey methods vary and accordingly are subject to the modifiable areal unit problem.

Our results show 11 distinct socio-economic clusters and provide a representative exemplar for each cluster. As previously discussed, purposive sampling has many advantages over random sampling, particularly for mixed-methods research. These exemplars are well suited to further in-depth studies and comparisons for research, especially mixed-methods and qualitative research. In these types of research designs, smaller number of cases to compare is not only more practical, but often more useful, for example when using methods such as focus groups, surveys, or participant observation. AP is also a powerful tool in overcoming issues with threshold-based values. For example, recent research has attempted to expand the concepts of rurality and urbanity for regions of varying spatial extents (Bibby & Brindley, 2014; Copus, 2015; Dicka et al.,
| Cluster # | Exemplar ZCTA | Population Density | % of Population 19 and under | % of Population 20–24 | % of Population 25–64 | % of Population 65 and above | % Hispanic or Latino | Average Family Size | % Below the Poverty Level | Per Capita Income (Dollars) | % Individuals Income | Management, Business, Science, and Arts Occupations | Service Occupations | Sales and Office Occupations | Natural Resources, Construction, and Maintenance Occupations | Production, Transportation, and Material Moving Occupations | Married (except separated) | Less than High School | High School Graduate (or equivalent) | Some College or Associate's degree |
|-----------|---------------|--------------------|-------------------------------|-----------------------|----------------------|-----------------------------|---------------------|--------------------|---------------------|-----------------------------|-------------------------|---------------------------------------------------------------------|-----------------------|------------------------|---------------------------------------------------------------|-----------------------------------------------|---------------------------------|------------------|-------------------------|--------------------------|
| 1         | 06040 Manchester, CT | 2.45 -0.17 | 22.30 -0.54 | 26.90 0.24 | 27.80 0.39 | 24.20 -0.22 | 27.10 0.27 | 27.20 0.29 | 26.80 0.22 | 26.60 0.02 | 31.30 0.98 | 18.70 0.49 | 46.60 4.01 |
| 2         | 12134 Northville, NY | 6.84 0.17 | 75.44 0.00 | 58.19 0.00 | 841.01 0.00 | 1040.10 0.00 | 4393.65 0.00 | 8001.17 0.00 | 26.35 0.00 | 1720.23 0.00 | 2620.60 0.00 | 35.20 0.00 | 54.20 0.00 |
| 3         | 14505 Marion, NY | 10.50 0.10 | 75.44 0.00 | 58.19 0.00 | 841.01 0.00 | 1040.10 0.00 | 4393.65 0.00 | 8001.17 0.00 | 26.35 0.00 | 1720.23 0.00 | 2620.60 0.00 | 35.20 0.00 | 54.20 0.00 |
| 4         | 28352 Laurnburg, NC | 13.80 0.10 | 75.44 0.00 | 58.19 0.00 | 841.01 0.00 | 1040.10 0.00 | 4393.65 0.00 | 8001.17 0.00 | 26.35 0.00 | 1720.23 0.00 | 2620.60 0.00 | 35.20 0.00 | 54.20 0.00 |
| 5         | 35713 Niangua, MO | 15.20 0.10 | 75.44 0.00 | 58.19 0.00 | 841.01 0.00 | 1040.10 0.00 | 4393.65 0.00 | 8001.17 0.00 | 26.35 0.00 | 1720.23 0.00 | 2620.60 0.00 | 35.20 0.00 | 54.20 0.00 |
| 6         | 75093 Plano, TX | 33.40 0.10 | 75.44 0.00 | 58.19 0.00 | 841.01 0.00 | 1040.10 0.00 | 4393.65 0.00 | 8001.17 0.00 | 26.35 0.00 | 1720.23 0.00 | 2620.60 0.00 | 35.20 0.00 | 54.20 0.00 |
| 7         | 75496 Wolfe City, TX | 40.80 0.10 | 75.44 0.00 | 58.19 0.00 | 841.01 0.00 | 1040.10 0.00 | 4393.65 0.00 | 8001.17 0.00 | 26.35 0.00 | 1720.23 0.00 | 2620.60 0.00 | 35.20 0.00 | 54.20 0.00 |
| 8         | 80503 Longmont, CO | 40.80 0.10 | 75.44 0.00 | 58.19 0.00 | 841.01 0.00 | 1040.10 0.00 | 4393.65 0.00 | 8001.17 0.00 | 26.35 0.00 | 1720.23 0.00 | 2620.60 0.00 | 35.20 0.00 | 54.20 0.00 |
| 9         | 80209 Cheyenne, WY | 40.80 0.10 | 75.44 0.00 | 58.19 0.00 | 841.01 0.00 | 1040.10 0.00 | 4393.65 0.00 | 8001.17 0.00 | 26.35 0.00 | 1720.23 0.00 | 2620.60 0.00 | 35.20 0.00 | 54.20 0.00 |
| 10        | 80511 Phoenix, AZ | 40.80 0.10 | 75.44 0.00 | 58.19 0.00 | 841.01 0.00 | 1040.10 0.00 | 4393.65 0.00 | 8001.17 0.00 | 26.35 0.00 | 1720.23 0.00 | 2620.60 0.00 | 35.20 0.00 | 54.20 0.00 |
| 11        | 92570 Parris, CA | 40.80 0.10 | 75.44 0.00 | 58.19 0.00 | 841.01 0.00 | 1040.10 0.00 | 4393.65 0.00 | 8001.17 0.00 | 26.35 0.00 | 1720.23 0.00 | 2620.60 0.00 | 35.20 0.00 | 54.20 0.00 |
| Place Name       | Exemplar ZCTA |
|------------------|---------------|
| Manchester, CT   | 06040         |
| Northville, NY   | 12134         |
| Marion, NY       | 14505         |
| Laurinburg, NC   | 28352         |
| Niangua, MO      | 65713         |
| Plano, TX        | 75093         |
| Wolfe City, TX   | 75496         |
| Longmont, CO     | 80503         |
| Chyenne, WY      | 82009         |
| Phoenix, AZ      | 85051         |
| Perris, CA       | 92570         |

| Total Population | 35306 |
| Area Square Miles | 6.20  |
| Population per Square Mile | 5695.49 |

| % Male Population | 48.2  |
| % White          | 74.1  |
| % Black or African American | 15.7  |
| % American Indian and Alaska Native | 1.1   |
| % Asian          | 6.5   |
| % Native Hawaiian and Other Pacific Islander | 0.1   |

| % Other Race | 6.4   |
| Place of Birth - Native | 85.7  |
| Foreign Born - Europe | 16.9  |
| Foreign Born - Asia | 36.3  |
| Foreign Born - Africa | 5.3   |
| Foreign Born - Oceania | 0.2   |
| Foreign Born - Latin | 38.4  |

| Foreign Born - Northern America | 2.9   |
| LSAH - English Only | 80.0  |
| LSAH - Spanish | 9.7   |
| LSAH - Other Indo-European Languages | 6.9   |
| LSAH - Asian and Pacific Islander Languages | 2.4   |
| LSAH - Other Languages | 1.0   |

Bachelor's degree: 19.8%, 0.62, 11.20, 0.31, 10.30, 0.38, 5.60, 0.88, 39.50, 2.69, 10.30, 0.38, 32.90, 2.00, 20.00, 0.64, 10.00, 0.41, 12.70, 0.82
Graduate or Professional degree: 13.5%, 0.71, 6.60, 0.14, 8.70, 0.11, 5.40, 0.29, 24.50, 2.07, 6.30, 0.18, 25.40, 2.18, 12.80, 0.62, 4.60, 0.39, 14.20, 0.64
Table 4. Summary of the distinct features of each cluster.

| #  | ZCTA        | Location     | Distinctive Characteristics                                                                 |
|----|-------------|--------------|---------------------------------------------------------------------------------------------|
| 1  | 06040       | Manchester,  | Overall close to average in all categories. Slightly higher than average population of 20–24 and 25–64 year olds. Slightly higher than average Per Capita Income and percent of population with income greater than or equal to $200,000. Higher than average jobs in Management, Business, Science, and Arts Occupations. Higher than average Hispanic or Latino population and percent of population with high percent of Native born residents and higher than average foreign born Latin American population. |
| 2  | 12134       | Northville,  | Exemplar with lowest percent of population 19 years and under. Extremely higher than average Hispanic or Latino population and percent of population with high percent of Native born residents and higher than average foreign born Latin American population. |
| 3  | 14505       | Marion, NY   | Exemplar with lowest percent of population with income greater than or equal to $200,000; however, still has an average Per Capita Income. Slightly higher than average White population. Exemplar with the lowest American Indian and Alaska Native population. |
| 4  | 28352       | Laurenburg,  | Higher than average percent of population below the poverty level. Lowest married population of all exemplars. Lower than average White Population, and significantly higher than average Black or African American population and American Indian and Alaska Native population. Extremely higher than average Hispanic or Latino population. Extremely higher than average family size. Exemplar with highest percent of Service Occupations and Sales and Office Occupations. Lower than average Spanish speaking population. |
| 5  | 67513       | Niangua, MO  | Average age distribution. Highest exemplar for percent of population with only a High School Graduate. Slightly higher than average white population and slightly lower than average Black or African American population. Exemplar with highest percent race identity distribution. |
| 6  | 75093       | Plano, TX    | Exemplar with lowest percent of population 20–24 years old. Higher than average percent of population 25–64 years old. Extremely wealthy exemplar with highest Per Capita Income and percent of population with income greater than or equal to $200,000. Higher than average jobs in Management, Business, Science, and Arts Occupations, with significantly lower number of jobs in other occupation categories. Extremely higher than average percent of population with a Bachelor’s degree and Graduate or Professional degree and lack of under educated population. Extremely higher than average Asian population, foreign born Asia population, and Asian and Pacific Islander speaking population. |
| 7  | 75496       | Wolfe City,  | Most average exemplar with very few categories above or below 0.05 standard deviations from the mean. Slightly higher foreign born Latin America population. |
| 8  | 80503       | Longmont, CO | Higher than average population of 20–24 year olds. Significantly higher than average Per Capita Income and percent of population with income greater than or equal to $200,000. High number of jobs in Management, Business, Science, and Arts Occupations. Extremely higher than average percent of population with a Bachelor’s degree or Graduate or Professional degree. Fairly average race identity distribution. |
| 9  | 82009       | Cheyenne, WY | An average exemplar with a slightly higher than average Per Capita Income. The exemplar with the highest married population. Slightly higher than average percent of population with a Bachelor’s degree or Graduate or Professional degree. |
| 10 | 85051       | Phoenix, AZ  | High percent of population 19 years and under. Higher than average Hispanic or Latino population. Higher than average family size. Exemplar with highest percent of Service Occupations and Sales and Office Occupations. Lower than average married population. Lower than average white population with a higher than average percent race identification of Some Other Race. Significantly lower percent of Native born residents and higher than average foreign born Latin American. Higher than average Spanish speaking population. |
| 11 | 92570       | Perris, CA   | Higher than average percent of population 19 years and under, and significantly lower than average percent of population 65 years and over. Extremely higher than average Hispanic or Latino population. Extremely higher than average family size. Higher than average percent of population below the poverty level and lowest exemplar for Per Capita Income. Percent of population with less than a high school degree. Lower than average White population and extremely high percent race identification of Some Other Race. Lower percent of Native born residents and higher than average foreign born Latin American. Extremely higher than average Spanish speaking population. |

2019; Nelson & Rae, 2016; Van Eupen et al., 2012; Waldorf, 2006), often attempting to overcome the unrealistic clear dichotomy created by these terms. Used within those contexts, AP can provide a data-defined basis for determining new values for rural/urban indexes, investigating the major factor differentiating (equating) localities at various scales, or re-shaping interaction areas, like regional urban systems (Burger et al., 2014), based on multiple metrics. Affinity propagation can further help in identifying study areas for comparative studies, highlighting the areas upon which similarities could be investigated, an invaluable asset in today’s polycentric urban analytical frameworks (Peck, 2015). The exemplars and clusters can also be used to select site locations for experimental design where the socio-economic factors included in the affinity propagation analysis need to be controlled.

Spatially, some of the exemplars seemed geographically removed from regions with the highest density of a given cluster. For example, Cluster 9’s exemplar is Longmont, CO, but the point density maps show high density of cluster 9 in the Boston–Washington corridor. Data-wise, the exemplars are the most representative ZCTA for a given cluster and spatial location was not included in the input data for clustering (i.e. we did not consider if a given ZCTA was similar to its neighbors and the exemplar selection did not consider if the exemplar was located in a region with a high concentration of that cluster). The potential role of spatial autocorrelation in the clustering and selection of exemplars is a potentially interesting area of future research.

While enhancing well-established geographic patterns at national and local scales, the AP analysis illuminates new relationships among communities at the zip code level. For example, higher percentages of Black residents are found in both the Deep South and key industrial centers of the Midwest and the Northeast, which uncovers some of the Great Migration’s lasting effects (Fliedstein, 1981), thereby influencing the emergence of a new cluster. Another example is the higher percentages of Latino and Spanish-speaking residents in the Southwest due to colonial and post-colonial Hispanic migration patterns alike (Hudson, 2002; Mines, 1981; Rouse, 1991). Economically-defined regions like the ‘Corn Belt’ are also easily discernable and match remarkably well with spatial delineations of the region found in previous studies (Hart, 1967; Hudson, 2002). At metropolitan scales,
legacy cities’ urban cores starkly contrast with their suburban counterpoints. In Michigan, for example, the inner-city ZCTAs of Detroit, Flint and Saginaw fall into cluster 4 and stand out as areas of high poverty, low white populations, high self-defined black populations, and lower educational attainment, while the surrounding suburbs are part of cluster 8 and are highlighted as areas of high educational attainment, higher income and higher white populations. The close proximity of cluster 3 and cluster 8 can also be seen in Chicago, Cleveland, Dallas, Buffalo, Raleigh and numerous others (Main Map). These urban/suburban patterns create ‘closer’ socio-demographic distance between physically far places.

The methodology provides a unique approach to clustering and exemplar selection, but the method used is computationally intensive, especially in terms of RAM requirements. The relationship between the number of units of analysis and RAM is exponential – somewhere between the 3rd and 4th power. This relationship limited the scale of analysis in this study to ZTCA rather than a finer level of analysis such as Census tracts. Using a Windows-based server with 128Gb of RAM, the ~32,000 ZCTA was the limit of our computational capabilities without using some sort of sampling approach. The AP algorithm is also sensitive to the preference value given. The preference value does not linearly correspond to the number of output clusters. Furthermore, even small changes in the preference value that results in the same number of clusters may result in slightly different clusters compositions and different exemplars. In this sense, AP can be seen as providing sets of potential exemplars for generating clusters within the data, rather than selecting a fixed sub-set of clusters. Further research is needed to investigate the impact of the preference value on exemplar and clustering results.

This study has demonstrated to potential applications of affinity propagation to typify socio-economic data and illustrate the results cartographically. This approach addresses major methodological and cartographic challenges of working with high dimensional socio-economic data. The resulting map illustrates interesting spatial socio-economic patterns as a result of complex socio-historical processes and help visualize the heterogeneous nature of the continental U.S. population. Applications of this approach are broad in the areas of market analysis and social research using mixed methods.

**Software**

Microsoft Excel was used to clean and format the census data. R with the packages FactoMineR (Le et al., 2008) and APCluster (Bodenhofer et al., 2011) were used to run the principle component analysis and
affinity propagation, respectively. Initial mapping and generalization was done using ArcGIS Pro 2.2 with final cartography done in Acrobat Photoshop CC.

Acknowledgements

The authors would like to thank Dr. David Patton at Central Michigan University for providing feedback on the manuscript, as well as the students in the Advanced Cartography class at Central Michigan University for their feedback on early drafts the map and John Nelson at ESRI for inspiration in the map aesthetics and the use of the desaturated ‘blue marble’ Earth imagery. We would like to thank the reviewers for their comments and useful feedback.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Department of Geography and Environmental Studies and the Center for Geographic Information Science at Central Michigan University.

References

Bibby, P., & Brindley, P. (2017). Rural-Urban classification of local authority districts in England: User Guide. Office of National Statistics.

Bodenhofer, U., Kothmeier, A., & Hochreiter, S. (2011). APCluster: An R package for affinity propagation clustering. Bioinformatics (oxford, England), 27(17), 2463–2464. https://doi.org/10.1093/bioinformatics/btr406

Bolton, T., & Hildret, P. (2013). Mid-sized cities: Their role in England’s economy. Centre for Cities. http://www.centreforcities.org/wp-content/uploads/2014/08/13-06-18-Mid-Sized-Cities.pdf

Brewer, C. A. (2019). http://www.ColorBrewer.org, Accessed 1 May 2019

Brookfield, K., Gray, T., & Hatchard, J. (2005). The concept of fisheries-dependent communities - A comparative analysis of four UK case studies: Shetland, Peterhead, north shields and Lowestoft.

Bureau, U. S. C. (2010). American factfinder, 2010 census. Sex by Age. http://factfinder2.census.gov

Bureau, U. S. C. (2011). 2007–2011 American community survey. Household Type (Including Living Alone). http://factfinder2.census.gov

Burger, M. J., Meijers, E. J., & Van Oort, F. G. (2014). The development and functioning of regional urban systems. Regional Studies, 48(12), 1921–1925. https://doi.org/10.1080/00343404.2014.979782

Cardille, J. A., & Lambois, M. (2010). From the redwood forest to the Gulf Stream waters: Human signature nearly ubiquitous in representative US landscapes. Frontiers in Ecology and the Environment, 8(3), 130–134. https://doi.org/10.1890/080132

Casellas, A., & Galley, C. C. (1999). Regional definitions in the European Union: A question of disparities? Regional Studies, 33(6), 551–558. https://doi.org/10.1080/0034309950078242

Chinni, D., & Gimpel, J. (2010). Our patchwork nation: The surprising truth about the “real” America. Gotham Books.

Copus, A. K. (2015). The new rural Economy and macroscale patterns. Territorial Cohesion in Rural Europe: the Relational Turn in Rural Development, 76, 11–34.

Culcasi, K. (2010). Constructing and naturalizing the Middle East. Geographical Review, 100(4), 583–597. https://doi.org/10.1111/j.1931-0846.2010.00059.x

Dicka, J. N., Gessert, A., & Sinicak, I. (2019). Rural and non-rural municipalities in the Slovak Republic. Journal of Maps, 15(1), 84–93. https://doi.org/10.1080/17445647.2019.1615010

Dueck, D., Frey, B. J., & Ilee. (2007). Non-metric affinity propagation for unsupervised image categorization [Paper presentation]. 11th IEEE International Conference on Computer Vision, Rio de Janeiro, BRAZIL.

Fliedstein, N. (1981). Going north, migration of blacks and whites from the South, 1900–1950. Academic Press.

Frey, B. (2009). Affinity Propagation FAQ. http://www.psi.toronto.edu/affinitypropagation/faq.html

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. Science, 317, 972–976.

Gibson, C., & Jung, K. (2002). Historical census statistics on population totals by race, 1790 TO 1990, and By Hispanic Origin, 1970 TO 1990, for the United States, regions, divisions, and states. United States Census Bureau. https://www.census.gov/content/dam/Census/library/working-papers/2002/demo/POP-wp0056.pdf

Goss, J. (1995). We know who you are and we know where you live: The instrumental rationality of geodemographic systems. Economic Geography, 71(2), 171–198. https://doi.org/10.2307/144357

Green, P. E., Frank, R. E., & Robinson, P. J. (1967). Cluster analysis in test market selection. Management Science, 13(8), B-387–B-400. https://doi.org/10.1287/mnsc.13.8.B387

Greenacre, M. (2005). Weighted metric multidimensional scaling. New Developments in Classification and Data Analysis, 141–149. https://doi.org/10.1007/3-540-27373-5_17

Hart, J. (1967). Change in the Corn Belt. Geographical Review, 76(1), 51–72. https://doi.org/10.2307/214784

Hazzi, N. A., Moreno, J. S., Ortiz-Molv, C., & Palacio, R. D. (2018). Biogeographic regions and events of isolation and diversification of the Endemic Biota of the Tropical Andes. Proceedings of the National Academy of Sciences, 115(31), 7985–7990. https://doi.org/10.1073/pnas.1803908115

Hedlund, M. (2016). Mapping the socioeconomic landscape of rural Sweden: Towards a typology of rural areas. Regional Studies, 50(3), 460–474. https://doi.org/10.1080/00343404.2014.924618

Hudson, J. (2002). Across this land: A regional geography of the United States and Canada. Johns Hopkins.

Kopanakis, I., & Theodoulidis, B. (2003). Visual data mining modeling techniques for the visualization of mining outcomes. Journal of Visual Languages and Computing, 14(6), 543–589. https://doi.org/10.1016/j.jvlc.2003.06.002

Le, S., Josse, J., & Husson, F. (2008). Factominer: An R package for multivariate analysis. Journal of Statistical Software, 25(1), 1–18.

Liebich, M. (2008). A region of hope, a region of despair: Print media, geographical imagination, and the gogebic iron range mining boom. Historical Geography, 36(1), 182–207.

Lynd, R. S. (1929). Middletown: A study in American Culture. Harcourt, Brace, & Co.
