[AM-02-009] Classification and Clustering

Abstract

Classification and clustering are often confused with each other, or used interchangeably. Clustering and classification are distinguished by whether the number and type of classes are known beforehand (classification), or if they are learned from the data (clustering). The overarching goal of classification and clustering is to place observations into groups that share similar characteristics while maximizing the separation of the groups that are dissimilar to each other. Clusters are found in environmental and social applications, and classification is a common way of organizing information. Both are used in many areas of GIS including spatial cluster detection, remote sensing classification, cartography, and spatial analysis. Cartographic classification methods present a simplified way to examine some classification and clustering methods, and these will be explored in more depth with example applications.

Keywords: basic analytical methods, classification, clustering, distance, spatial patterns

Author & citation

Lamb, D. (2020). Classification and Clustering. The Geographic Information Science & Technology Body of Knowledge (1st Quarter 2020 Edition), John P. Wilson (ed.).
DOI: 10.22224/gistbok/2020.1.11.

An earlier version can also be found at:

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Spatial Cluster Analysis. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers.

Explanation

1. Overview
2. Univariate Classification and Clustering
3. Multivariate Classification and Clustering

1. Overview

Classification and clustering are often confused with each other, or used interchangeably. Their definitions changing slightly depending on the discipline or sub-discipline. In either case, the goal is to generalize detailed information contained in attributes into a smaller number of classes (categories or groups). If an observation is part of a category, it is said to be a member of that group. Membership in a category means an observation cannot be a member of any other category, or the categories are said to be mutually exclusive. That is, there is no overlap between the boundaries of each class.

Clustering and classification are distinguished by whether the number and type of classes are known beforehand (classification), or if they are learned from the data (clustering). This

is sometimes distinguished as supervised learning (classification), and unsupervised learning (clustering). Geographic location may or may not be incorporated into either approach.

An example of pre-defined categories used in remote sensing classification are land cover classes, such as Water or Barren Land. There is an existing number of categories for land cover (potentially several hundred different categories). Each of these land cover classes has certain characteristics associated with it (color, reflectance, etc…). This information can be used to place new observations into these classes. Many of these classification methods learn to differentiate between classes based on a training dataset where an observation's class membership and attributes are defined. Many classification methods are probabilistic in nature, meaning they estimate the probability of being a member of a particular group.

Clustering attempts to create categories based on the similarities between observations' attributes; more similar observations are placed in the same group together. Sometimes clustering methods will attempt to determine the number of groups, and other times the analyst or researcher will need to provide this information. Spatial clustering examines the distribution of spatial features, and non-spatial clustering relies on characteristics of observations to group them. Spatial and non-spatial may be combined in different methods.

As can be seen in Table 1, classification and clustering approaches touch many different areas of GIS&T. Since there is topic has such a large scope, this section will focus on univariate cartographic classification. This will provide a general overview of how data is placed into different categories that might be extrapolated to more complex applications. Before moving to those examples, the underlying concept of similarity should be discussed.

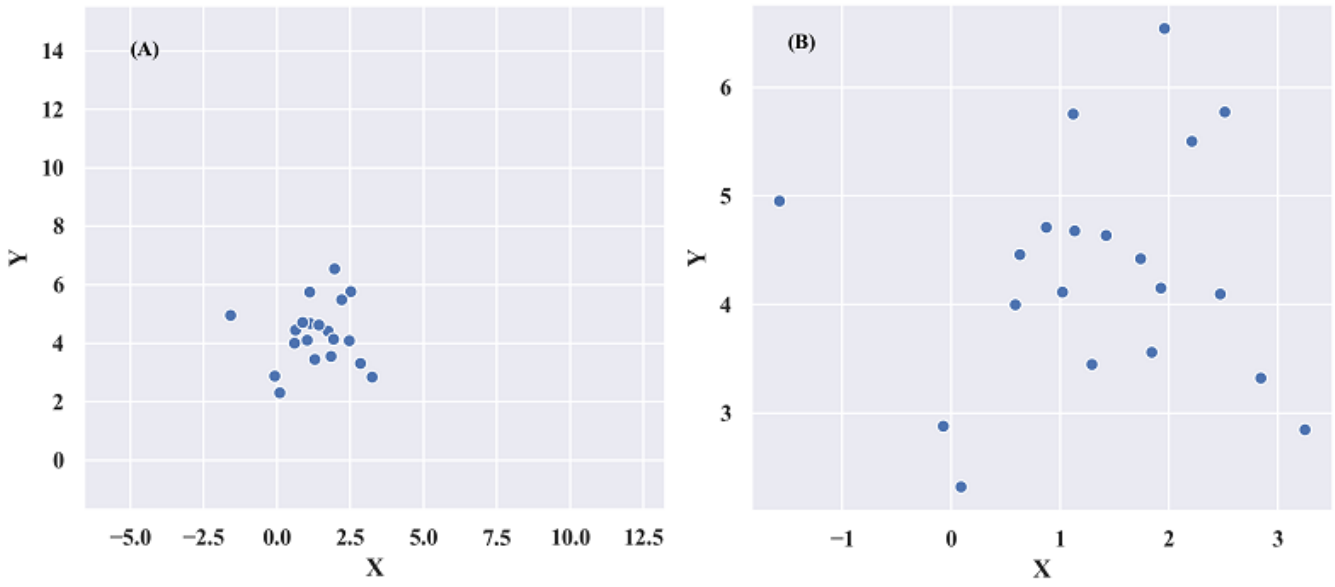
Table 1. Cluster and Classification Techniques in Different Geospatial Areas

Area	Purpose / Description	Example Methods
Spatial Statistics (Bailey & Gatrell, 1995; O'Sullivan & Unwin, 2010)	In point pattern analysis, methods are used to identify the presence of spatial clusters using distance between points. Other methods will use spatial autocorrelation to detect cluster patterns in the data.	K-Function, Kernel Density Analysis, Quadrat Analysis, Average Nearest Neighbor, or Moran's I
Remote Sensing (Lu & Weng, 2007; Smith, Goodchild, & Longley, 2007)	Performed using multi-band imagery and remotely sensed data. Each observation is a pixel or cell in the study area, and the characteristics of that pixel (e.g. the spectral signature) is used to place it in a class. Classification is usually called supervised classification, and clustering is called unsupervised classification.	Supervised classification: Logistic Regression, Support Vector Machines, or Random Forest Classifier Unsupervised classification: K-means clustering
Spatial Data Mining (Miller & Han, 2009)	Methods to detect patterns in data include many variables, sometimes also uses location information. Classification is called supervised learning, and clustering is called unsupervised learning.	Supervised learning: Logistic Regression, Support Vector Machines, or Random Forest Classifier. Unsupervised learning: K-means, hierarchical clustering, or density-based (e.g., DBSCAN)

Area	Purpose / Description	Example Methods
Geodemographics (Alexiou & Singleton, 2015)	Clustering is used to discover groups within the population, or what are called population segments (e.g. Commuters with Young Families) that are tied to a geographic area (typically a census unit).	Segmentation: K-means clustering
Statistical Clustering	The purpose for this clustering may be to reduce the complexity of the data by combining variables into a single variable (such as through Principal Component Analysis). This area overlaps with the above areas as well.	Reduction techniques: multidimensional scacling, factor analysis, principal components analysis, K-means clustering, or lineara discriminant analysis

2.1 Applications of Clustering and Classification

It may be tempting to identify clusters visually, but this can be misleading. Consider Figure 1 which presents the same point data in different scales. With Figure 1A (left) the distribution of the points might be considered clustered, but when the scale is changed in Figure 1B (right), the distribution might be considered dispersed, or even random. How the boundary of a study area is defined can influence how spatial clusters are defined, and this is called edge effects. The methods described in Table 1 use statistical procedures to measure the degree of clustering in many different types of data. This avoids the problem of relying on a visual interpretation.



Figures 1a and 1b. The same distribution of points viewed at different scales; (A) appears to be clustered, and (B) appears to be disperse. Source: author.

Some spatial clustering relies on the x and y coordinates of points to determine whether clustering is present, or where the clusters are. data may be able to use the x and y coordinates of the points and distance between them to identify the clusters. More complex cluster detection requires understanding the spatial relationships between features or

phenomena, typically with polygons. Usually this relationship is described through a neighborhood graph or matrix that will tell the clustering method which features are neighbors or not. These relationships and distances can be compared to a theoretical random distribution to tell the degree of clustering (e.g. the Poisson distribution), or a different type of metric (Silhouette score).

Spatial clustering may include non-spatial attributes or variables. Non-spatial clustering will rely entirely on attributes of the observed data, but use ideas already familiar to GIS users such as Euclidean distances (see below). Clustering is seen in many real-world phenomena. The concept of Agglomeration in geography is the idea that similar businesses will be located near one another in order to share resources or customers (e.g. car dealerships). Another example, Similar types of crimes tend to be located near each other. Epidemiologists may be interested in where there are groups of infected individuals that are near each other in space and time that are not normal. These clusters may point to a source of exposure, or some unknown reason for their illness.

Similarly, classification methods might be used to predict if someone has a particular disease. Different diseases are treated as separate classes, and many variables (e.g. height, weight, age, career, etc…) are used to predict which class a person would be in. Classification is used in GIS and Cartography to develop thematic maps the group similar types of data together and color coded (explored in more depth below). Finally, both classification and clustering are used in remote sensing applications to group similar raster cells into homogeneous areas as in land cover analysis, creating categories like Deciduous Forests, Water, or Barren Land.

2.2 Similarity and Distance

The overarching goal of classification and clustering is to place observations into groups that share similar characteristics while maximizing the separation of the groups that are dissimilar to each other. This naturally leads to the question: how are two observations similar (or dissimilar)? One approach is to use a distance metric that can be interpreted as a measure of similarity between pairs of observations. The shorter the distance the more similar the two will be. There are many ways to calculate the distance, but one often used in GIS&T is Euclidean distance. Given the location of two different observations or points each with a coordinate pair xi,yi , the distance can be calculated as in equation 1.

$$\sqrt{(((x_1 - x_2\ )^2 + (y_1 - y_2\ )^2\ )\ )}$$

(Equation 1)

If we expand the view of what a location is, it can incorporate not just a physical geographic location, but any a place in any numeric variable. In this view, Euclidean distance is calculated between any number of variables. Another common distance metric is Manhattan distance, shown in equation 2, and there are many others.

$$|x_1 - x_2| + |y_1 - y_2| \qquad \text{(Equation 2)}$$

There are other common approaches that may or may not use distance as a measure of similarity or proximity. A comparison of some of the major spatial clustering and

classification algorithms is presented in Table 2.

Table 2. Common Clustering and Classification Techniques and Their Approaches

Example Algorithm or Method	Use	Approach
Quadrat Analysis	Measurement of spatial clustering of points	Segments the study area into a grid, then counts the number of points in each cell
K-Function	Presence of spatial clustering of points	Multiple lags or distances are used from each point (imagine ripples in a pond from a handful of pebbles thrown in at once), and points are counted within eac ring. This is compared to a simulation of random points within the same area (Monte Carlo simulation).
Moran's I	Combines spatial information with an attribute	Similar to more traditional statistical methods and hypothesis testing.
K-Means	Spatial or non-spatial clustering	Relies on a measure of similarity to detect which data belong to which cluster. Explored in more depth below in this entry.
DBSCAN	Non-spatial clustering (but could use if for spatial)	Similar to a k-function, it looks for points within a distance. It also tries to find “noise” or random points that do not fall within clusters. It requires setting parameters like the minimum number of points to be considered for a cluster. These parameters can be difficult to set, and will change the outcome.
Random Forest	Classification	A popular artificial intelligence classification technique. Uses decision trees to identify the most important variables that sort data into different classes.

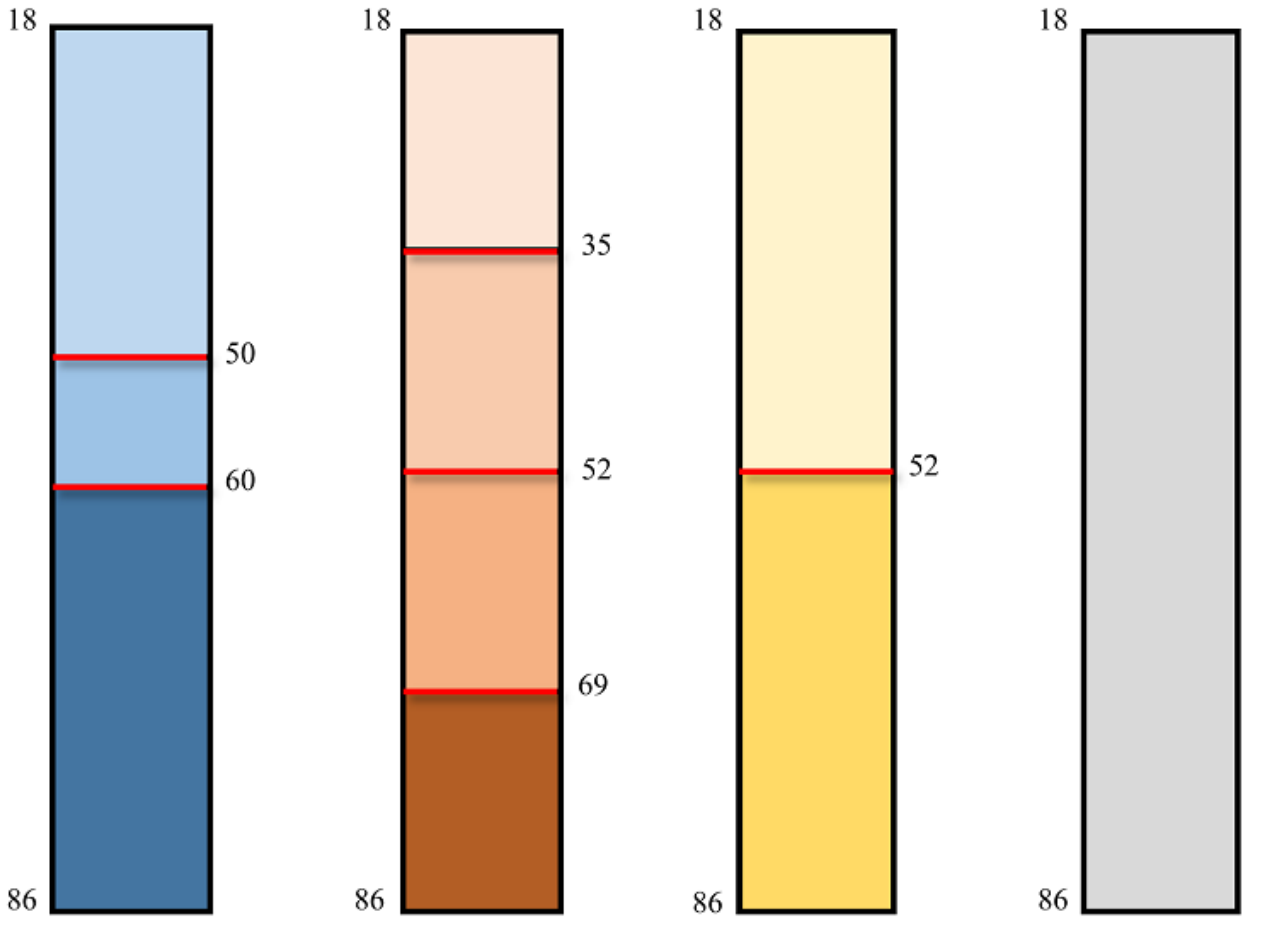
3. Univariate Classification and Clustering

Cartographic classification methods present a simplified way to examine some classification and clustering methods. Within cartography, classification is a process of simplifying or aggregating information into groups to be displayed on a map. Table 3 presents some of the most common approaches used in cartography. To distinguish these classes, every member of a group is assigned the same map symbol to their geographic information. As an example, for this type of mapping, univariate values are taken from a polygonal geographic unit like the United States (U.S.) Census Tract, and a color is assigned to each category (choropleth mapping). The cartographer selects the number of categories for the map. Then, the classification method selected will determine the boundaries of the classes (Figure 2). The boundaries define the lower/starting and upper/ending values for each group, sometimes these values are called ‘breaks’ (Brewer & Pickle, 2002). For additional information, see Statistical Mapping (Enumeration, Normalization, Classification).

Table 3. Common Classification Methods Used in Cartography and Choropleth Mapping

Type of Classification	Description
Unique values	Each value is its own class or group, and each group is assigned a color. Typically reserved for categorical data (e.g. nominal level data).
Manual Classification	The cartographer designates the bounds of each class as mutually exclusive groups.

Type of Classification	Description
Equal Interval	Uses the range from the variable and divides this by the number of classes, creating an interval.
Defined Interval	With this method the interval is selected first, and the number of classes derived from how many intervals are needed to cover the range.
Quantile or Percentile	Uses the percentage of values that fall in particular ranges, based on the number of classes selected. The same number of observations will exist in each category.
Natural Breaks / Jenks	An algorithmic approach to identify “natural” break points in the data (Jenks & Caspall, 1971). It is similar to the K-means clustering approach.
Standard Deviation	A statistical approach using the mean of the data, and the standard deviation. Often used to show extreme values or deviations from the mean (a diverging pattern in the symbology).
Equal Area	This approach uses the area of the polygons to determine the class breaks so that each group contains an equal proportion of the overall area (Brewer & Pickle, 2002; Lloyd & Steinke, 1977). An alternative to normalizing the variable by the polygon’s area.
Head/tail Breaks	Relatively new technique that is designed for variables with a skewed distribution (heavy tailed) (Jiang, 2013).



(A) Manual Classification with 2 class breaks

(B) Equal Interval Classification with 4 class breaks

(C) Defined Interval Classification with 2 class breaks

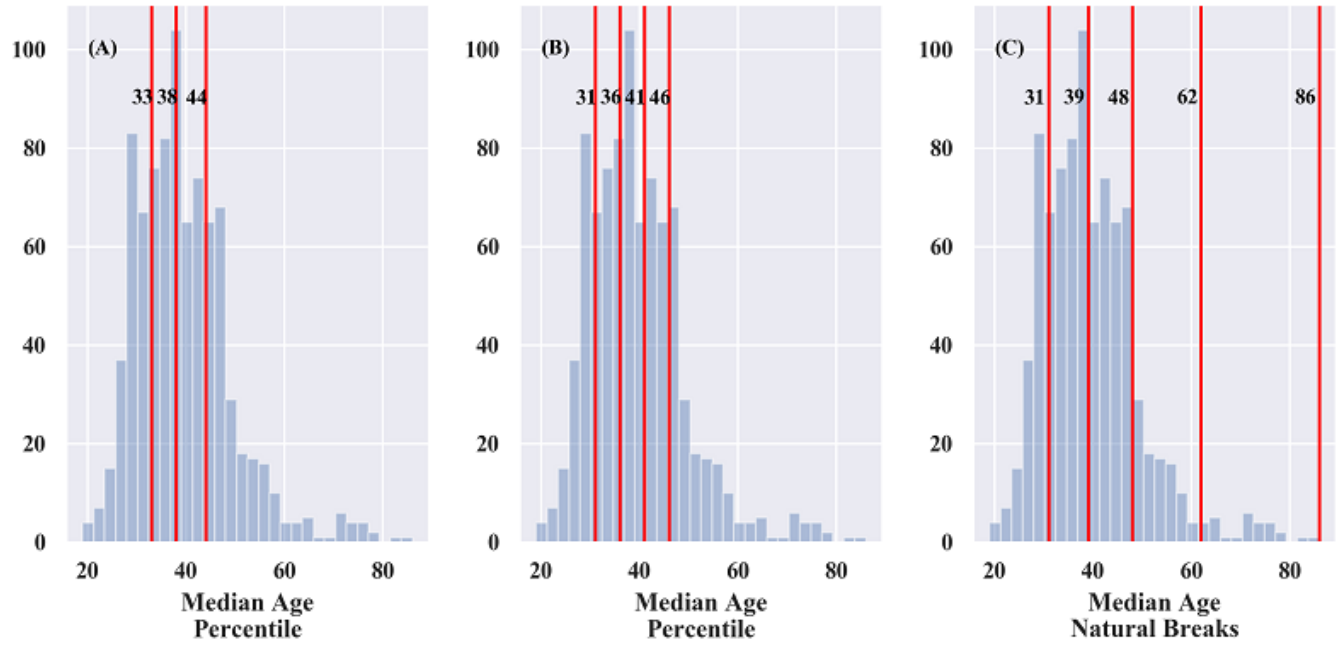
(D) Data range with no class breaks

Figures 2A - 2D. Data classification methods for grouping data using (A) manual classification, (B) equal interval classification, (C) defined interval, and applied to median age data (D). Source: author.

The most basic case is the unique value approach where each value is assigned a unique symbol, creating a class or category for each value. While it is possible to do this for numeric data, it is usually reserved for categorical (nominal level) data, otherwise there could be many different unique classes. Manual classification allows the cartographer to define the upper and lower limits of the boundaries, or the break points. Figure 2A presents two “breaks” at 50 and 60, creating three classes. Because these upper and lower limits are mutually exclusive, the first class runs from 18 to 50, the second class runs from 51 (or possibly 50.000001) to 60, and the third group runs from 61 (or possibly 60.0000001) to 86. Each group is assigned a color value, and in the case of sequential data, the colors progress from light to dark.

Equal Interval and Defined Interval present a simple classification scenario based on the characteristics of the data. Equal Interval sets the number of classes and the data range is divided by this number (e.g. (86-18)/5=17), creating an interval of 17. The values are placed in these known classes, as shown in Figure 2B. In the Defined Interval approach the cartographer selects the interval then the software determines the number of classes that will fit. In Figure 2C, the interval is 34 creating two classes.

Other approaches rely on the distribution of the variable. Frequency distributions are visualized as a histogram. Histograms divided the data into bins of equal widths (e.g. between 40 and 45), and count the number of values that fall inside each bin. This count is reflected in the bar height. Figure 3 shows a histogram for the median age variable. Percentiles are the percentage of the data that falls below the corresponding value. To define class breaks using percentile the number of classes is selected, then the range 0 to 100% is divided by this number. In Figure 3A, 3 classes result in percentiles at 25%, 50%, and 75%. These correspond to values of 33, 38, and 44 respectively. This means that 25% of the observations fall to the left of 33 on the histogram, 50% fall to the left of 38, and 75% fall to the left of 44. Keep in mind the height of the bar indicates the actual total number of observations. At 86, 100% of the observations fall to the left of the histogram. Figures 2B presents percentiles for 5 classes, but the idea remains the same.



Figures 3a - 3c. Frequency distribution of median age data and the breaks associated with (A) 3 class percentiles, (B) 5 class percentiles, and (C) 5 class Jenks' Natural Breaks method. Source: author.

Finally, Jenks' method is an iterative algorithmic approach that identifies 'natural' break points in the data (Jenks & Caspall, 1971). This method is closer to the idea of identifying clusters (groups) in the data, resulting in uneven intervals. The result of Jenks' method applied to the median age variable are shown in Figure 3C. The algorithm is complex, but a there is a similar approach called K-means clustering that can be demonstrated. K-means is also widely used in other areas. The k refers to the number of classes. There are approaches to help choose k, but these are not widely available in GIS packages (silhouette scores, or elbow plots).

Beginning with a smaller dataset, Figure 4 shows 20 observations along a number line. K-means clustering begins by selecting k, and three is used in this example. The algorithm begins the first iteration by generating three random values that fall within the variable's range. In Figure 5, there are three random values generated (a light blue, yellow, and dark grey point). These will serve as the center of each of the groups for a first iteration. Next, the algorithm calculates the distance from the original observation to each of the cluster centers. Figure 6 presents this distance as an arrow from the first observation on the left to each of the centers.

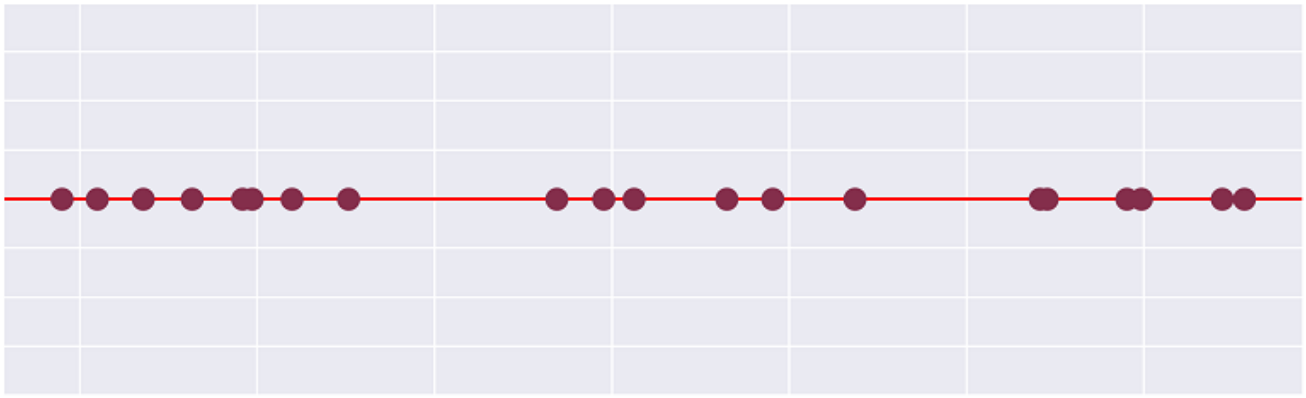


Figure 4. Twenty observations along a number line to demonstrate K-means algorithm. Source: author.

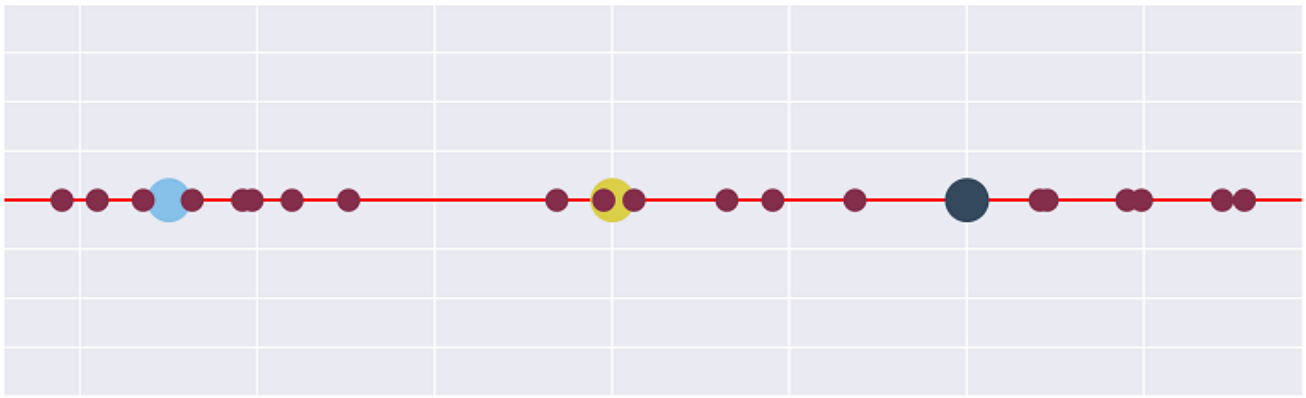


Figure 5. Twenty observations with three randomly generated cluster centers along the number line. This demonstrates the first step in the K-means algorithm. Source: author.

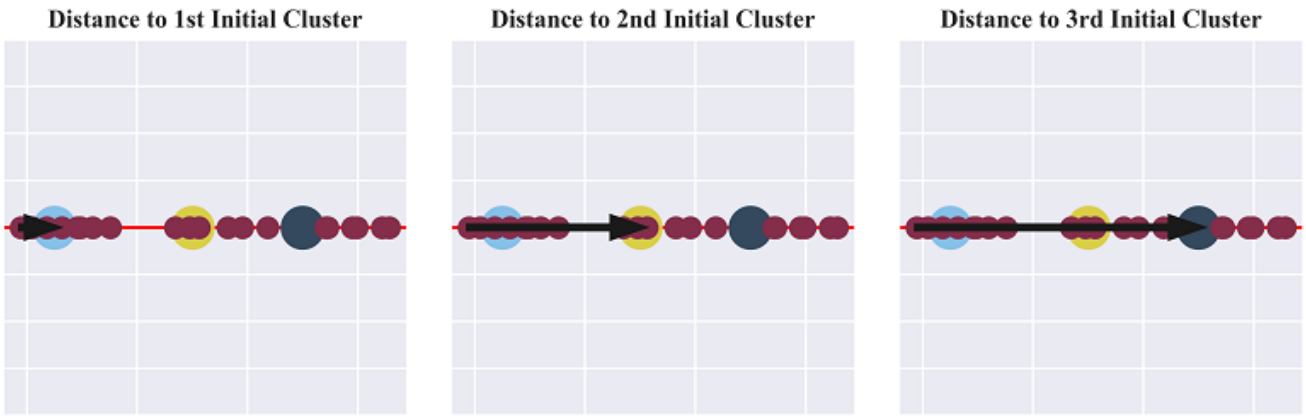


Figure 6. Calculation of the distance from the first observation on the left of the number line to each random cluster center. Source: author.

Next, the algorithm assigns each observation to the nearest cluster center based on the distance (Figure 7A). These create the first set of clusters, and the algorithm calculates the mean within each cluster (the vertical lines in Figure 7B represent the mean of each cluster). Now, the mean becomes the center of each cluster, and the distance is recalculated from each observation to these new centers. Again, observations are moved to different groups based on the distance. The mean for each of the clusters is calculated again, and the distance again, and so on. This process repeats until there are no changes to which cluster an observation belongs to. The algorithm begins a new iteration, creating random values for each cluster center. After many iterations, it will return the 'best' fitting clusters. It may turn out the first iteration was the best, but it will repeat the process as many times as the analyst will specify.

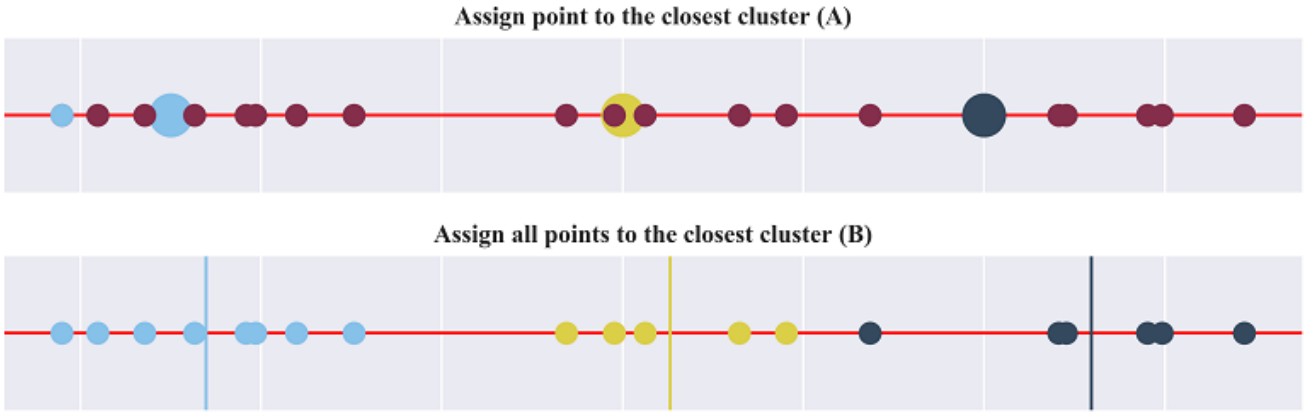


Figure 7a - 7b. During the K-means algorithm observations are assigned to their closest cluster center, (A) the leftmost observation is assigned to the closest cluster center (A), and (B) all observations are assigned to their closest cluster center and the mean of each group is calculated. Source: author.

At each iteration, the algorithm calculates a measure to determine how well those clusters fit the data. To do this, the algorithm uses the variance of each cluster and the total variance. The variance measures the distance of each group member to the mean of the cluster. The variance is a measure of the spread of the data. Figure 8 shows this spread as arrows, and each group has a different width. The total variance is the sum of the group variances, and the iteration that had the smallest total variance is returned as the best option. The result is class bounds that can have different interval widths, and unusual beginning and end points.

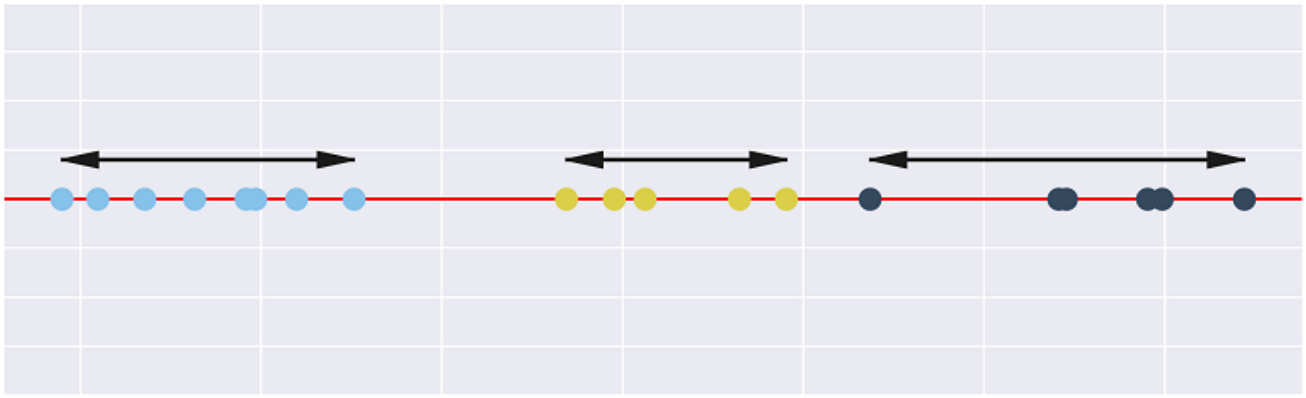


Figure 8. The K-means algorithm uses the variance of the different clusters (as demonstrated by the arrows) to determine the performance of these groupings. Source: author.

Which method should you use? Which method chosen and how many classes selected will have consequences on the map's final appearance and potentially the way the data are interpreted. This can be considered from both ethical and practical considerations (Harley, 1991; Monmonier, 1991). For comparison, Figure 9 presents the median age data for Hillsborough County, Florida using U.S. Census Tracts. Each tract is placed in a class depending on the median age for that tract. The class bounds change depending on the method, and the results can create very different interpretations of the underlying data. For example, Figure 9B shows class breaks using the Equal Interval method for 5 classes. This map creates the impression that most of the county falls in the 27 to 41 median age group. Whereas, the Natural Breaks method (Figure 9A) has varying class intervals (40 to 47 is small, compared to 62 to 86), and results in a more diverse county.

It can be difficult to choose which method to apply. Slocum et al. (2009, p. 68) provide some guidelines for mapping. Sometimes the shape of the frequency distribution (e.g. a normal distribution is appropriate for the percentile), or other characteristics of the data, might help. Monmonier (1991) suggests presenting the reader with a "dynamic sequence" of maps that show the extreme views of the data (Monmonier, 1991, p. 4).

Regardless of the method, one should take care to balance the interpretability of the classes, while letting the data speak for itself. In the case of the median age variable, using manually selected classes that reflect life stages (e.g. voting age in the United States is 18, or retirement age is 66) is logical and easily understood by the map reader. This implies some artistic license that is not always available or appropriate in other areas of classification and clustering.

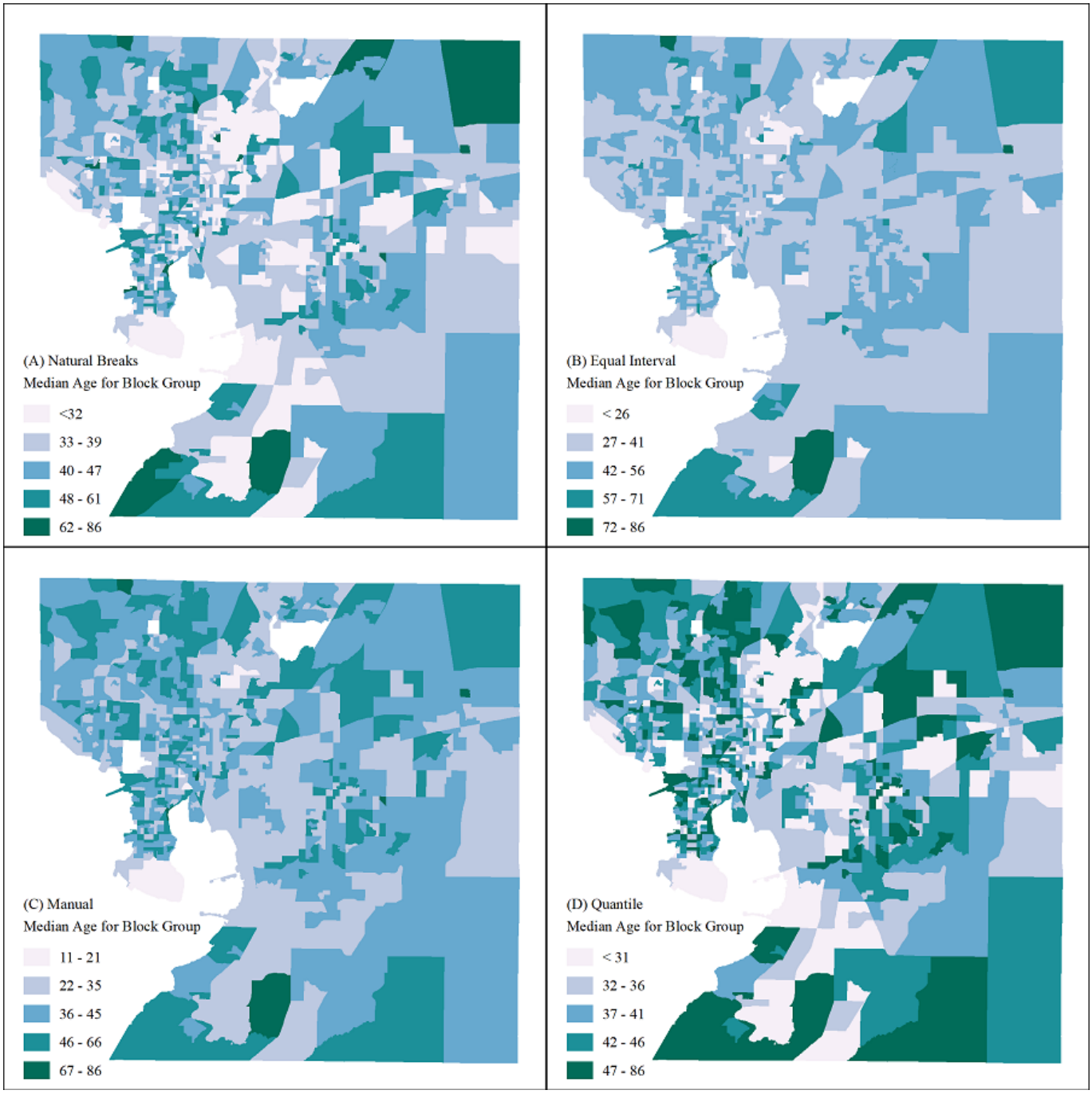


Figure 9A-9D. Comparison of cartographic classification methods using (A) Jenks' Natural Breaks, (B) Equal Interval, (C) Manual, and (D) Percentile / Quantile. Source: author.

4. Multivariate Classification and Clustering

Many classification and clustering methods are applied to multivariate data. The K-means approach can be expanded to include many attributes, and the algorithm remains the same. It still seeks the center of each cluster. Figure 10 shows clusters of observations from three variables in three-dimensions. The cluster that is shaded blue with a triangle shape overlaps with the green dot cluster along the first variable, but has very different results for the third variable (z-axis). It is the combination of characteristics that can create mutually exclusive clusters when dealing with multivariate data.

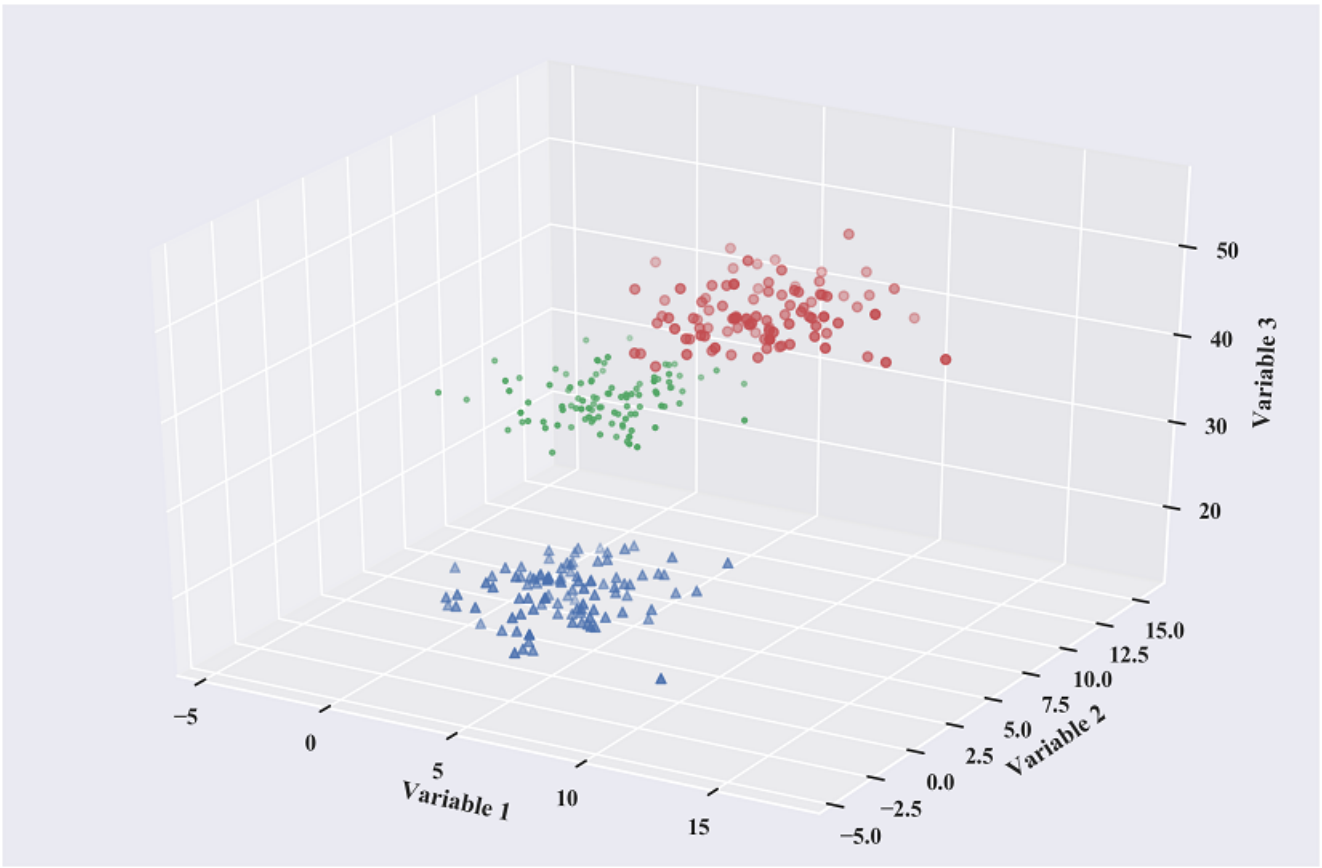


Figure 10. Multivariate clustering and classification attempts to separate groups based on more than one variable. Source: author.

References

Alexiou, A., & Singleton, A. D. (2015). Geodemographic Analysis. In C. Brundson & A. D. Singleton (Eds.), Geocomputation. A Practical Primer (pp. 137–151). London: Sage.

Bailey, T. C. and Gatrell, A. C. (1995) Interactive Spatial Data Analysis. Vol. 413, Longman Scientific & Technical, Essex.

Brewer, C. A., & Pickle, L. (2002). Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series. Annals of the Association of American Geographers, 92(4), 662–681.

de Smith, M. J., Goodchild, M. F., & Longley, P. A. (2007). Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools (2nd Edition). Troubador Publishing.

Harley, J. B. (1991). Can there be a cartographic ethics? Cartographic Perspectives, 10, 9-16.

Jenks, G. F., & Caspall, F. C. (1971). Error on Choroplethic Maps: Definition, Measurement, Reduction. Annals of the Association of American Geographers, 61(2), 217–244.

Jiang, B. (2013). Head/tail Breaks: A New Classification Scheme for Data with a Heavy-tailed Distribution. The Professional Geographer, 65, 482-494.

Lloyd, R., & Steinke, T. (1977). Visual and Statistical Comparison of Choropleth Maps. Annals of the Association of American Geographers, 67(3), 429–436.

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing, 28(5), 823–870.

Miller, H. J., & Han, J. (Eds.). (2009). Geographic Data Mining and Knowledge Discovery: An Overview. CRC Press, Taylor and Francis Group.

Monmonier, M. (1991). Ethics and Map Design: Six Strategies for Confronting the Traditional One-Map Solution. Cartographic Perspectives, 10, 3-8.

O'Sullivan, D. and Unwin, D. (2010) Geographic Information Analysis, 2nd Edition. John Wiley & Sons, Inc.

Slocum T. A., McMaster, R. B., Kessler, F. C., & Howard, H. H. (2009). Thematic Cartography and Geographic Visualization (3rd edition). Upper Saddle River, NJ: Pearson/Prentice Hall.