DATA CLASSIFICATION

Sandeep Talasila, GISP

INTRODUCTION

- Categorizing geographic features based on conditions
- Features in each class will have a common property and are displayed using identical symbology
- Manual or Automatic
- Defining number of classes is dependent on the data distribution and the question you are trying to answer

DATA MEASUREMENT

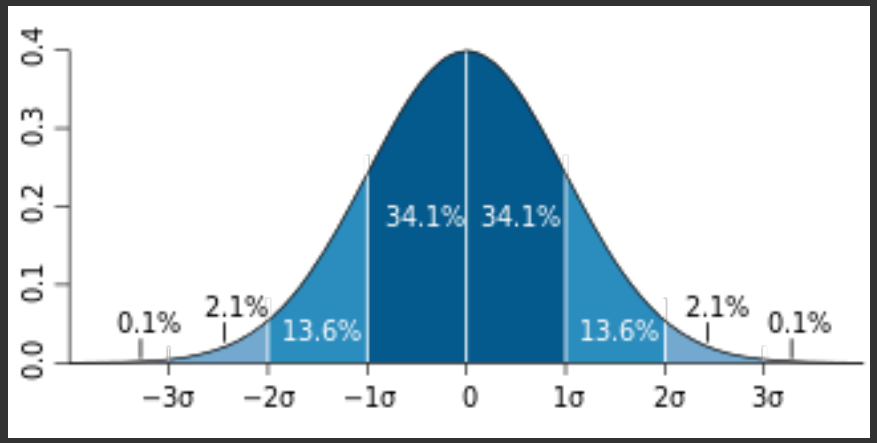
- Nominal
- Ordinal
- Interval
- Ratio

DERIVED DATA INDICES

- Ratio
  - Relation between two entities
- Proportion
  - Ratio of the number of items in one group (class) to the total items
- Percent
  - Proportions multiplied by 100
  - Percentage changes
- Rate
  - Similar to percentages except that the relationship is a value per a much larger value

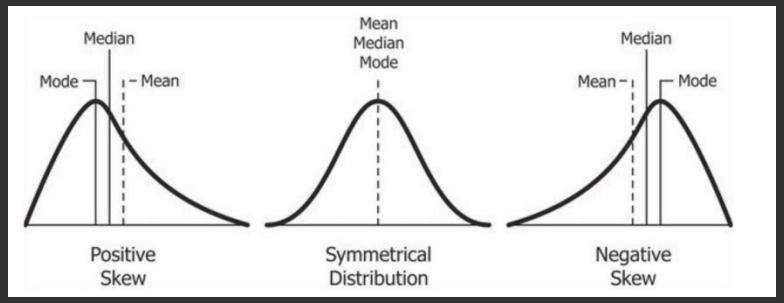
DESCRIPTIVE STATISTICS

- Sort – Rank Order
- Range
- Measures of Central Tendency
  - Mean, Median, Mode
- Measures of Dispersion
  - Variance
  - Standard Deviation
- Skewness
- Kurtosis



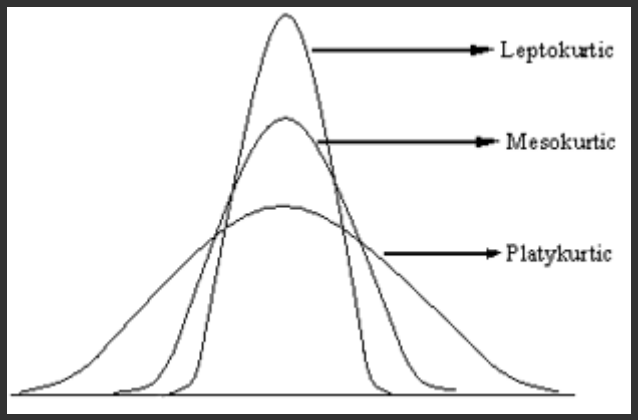
SKEWNESS AND KURTOSIS

Skewness = 0 for Normal Distribution



By Diva Jain - https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=84219892

Kurtosis = 3 for Normal Distribution

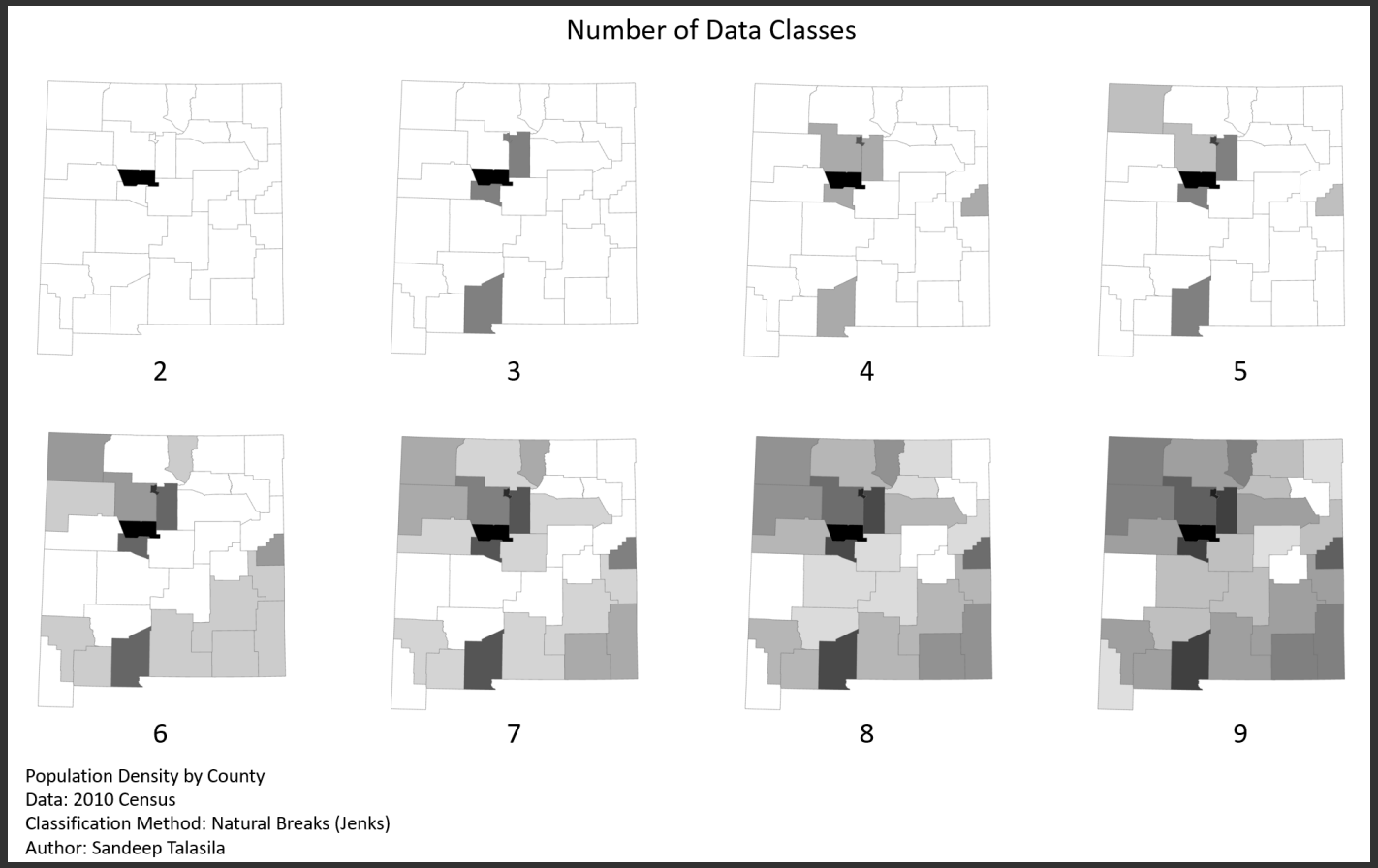


NUMBER OF DATA CLASSES

- C = 1 + 3.3 * log(n) – *Sturges (1926)*
  - n = number of observations

- Zero Observation Classes
  - Should categorize separately
  - Values with zero should be included for calculations (mean)
  - Values with No Data should not be included

NUMBER OF DATA CLASSES



DATA CLASSIFICATION REQUIREMENTS

- Contain full range of the data values
- Have neither overlapping values nor vacant classes
- Be enough in number for accuracy of data, but not numerous as to impute great degree of accuracy than necessary
- Divide data into reasonably equal groups of observations
- Have a logical mathematical relationship if practical

- Jenks and Coulson (1963)

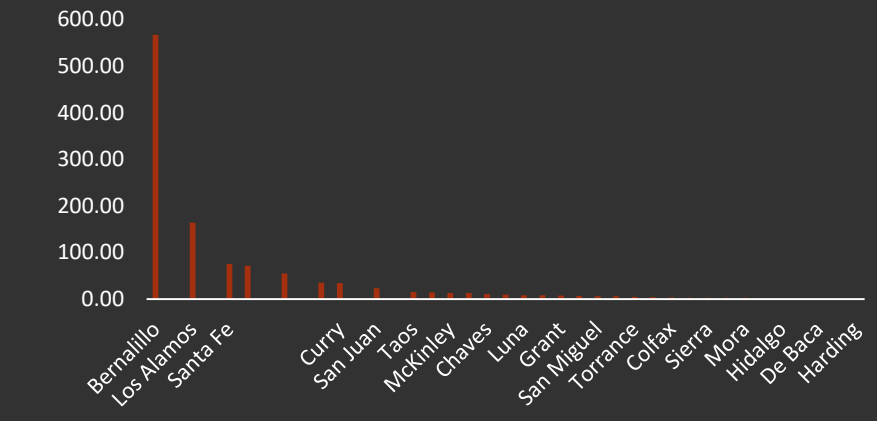
DATA CLASSIFICATION SCHEMES

- Natural Breaks
- Jenks Optimization
- Nested Means
- Mean and Standard Deviation
- Equal Interval
- Equal Frequency
- Arithmetic
- Geometric
- User Defined

NATURAL BREAKS

- Natural Breaks classes are based on natural groupings inherent in the data.
- Class breaks are identified that best group similar values and that maximize the differences between classes.
- Data-specific and not useful for comparing multiple maps built from different underlying information.

NATURAL BREAKS



Class	Min–Max	Frequency
1	164.33 – 567.69	1
2	75.48 – 164.32	1
3	54.86 – 75.47	2
4	35.43 – 54.85	1
5	23.46 – 35.42	2
6	14.97 – 23.45	1
7	0.33 – 14.96	25

County	Population Density	Gap Value	Gap Order	Class
Bernalillo	567.69			Class 1
		403.37	Gap 1	
Los Alamos	164.32			Class 2
		88.85	Gap 2	
Santa Fe	75.47			Class 3
Valencia	71.71	3.76		
		16.87	Gap 4	
Dona Ana	54.85			Class 4
		19.43	Gap 3	
Sandoval	35.42			Class 5
Curry	34.35	1.07		
		10.90	Gap 5	
San Juan	23.45			Class 6
		8.49	Gap 6	
Taos	14.96			Class 7
Lea	14.73	0.22		
McKinley	13.09	1.65		
Eddy	12.83	0.25		
Chaves	10.81	2.02		
Otero	9.63	1.18		
Luna	8.46	1.18		
Roosevelt	8.09	0.36		
Grant	7.43	0.67		
Rio Arriba	6.83	0.60		
San Miguel	6.20	0.62		
Cibola	5.98	0.22		
Torrance	4.90	1.09		
Lincoln	4.25	0.65		
Colfax	3.65	0.59		
Quay	3.14	0.51		
Sierra	2.83	0.31		
Socorro	2.69	0.14		
Mora	2.53	0.15		
Guadalupe	1.55	0.99		
Hidalgo	1.42	0.13		
Union	1.19	0.23		
De Baca	0.87	0.32		
Catron	0.54	0.33		
Harding	0.33	0.21		

NATURAL BREAKS

- Advantages
  - Considers distribution of data
  - Easily computed
- Disadvantages
  - Class breaks are subjective
  - Difficulty to determine breaks with large datasets
  - May miss natural spatial clusters

- Natural Breaks is good for data with an uneven distribution.

JENKS OPTIMIZATION

- Algorithm developed by Walter-Fisher (1958) and implemented by George Jenks (1977)
- Iterative process to determine smallest class variance
- Places similar data values in the same class by minimizing a measure of classification error.
- Goodness of variance Fit (GVF) ranges from 0 to 1
  - 0 – low accuracy
  - 1 – high accuracy

JENKS OPTIMIZATION

Data Distribution [4, 5, 9, 10]

Mean = 7

- **Step 1 – C**alculate the "sum of squared deviations for array mean" (SDAM).
  - SDAM = (4-7)^2 + (5-7)^2 + (9-7)^2 + (10-7)^2 = 9 + 4 + 4 + 9 = 26.

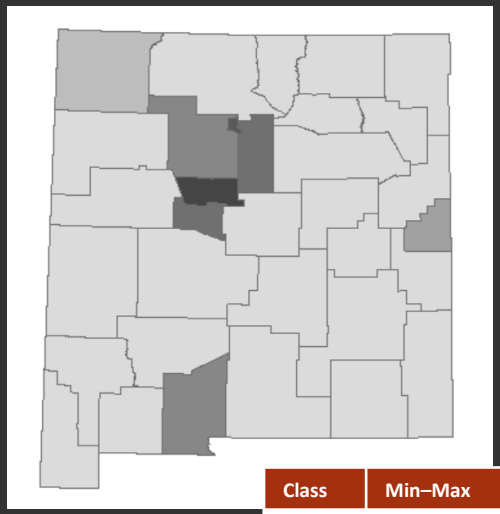
- **Step 2 –** For each range combination, calculate "sum of squared deviations for class means" (SDCM_ALL), and find the smallest one. SDCM_ALL is similar to SDAM, but uses class means and deviations.
  - For [4][5,9,10], SDCM_ALL = (4-4)^2 + (5-8)^2 + (9-8)^2 + (10-8)^2 = 0 + 9 + 1 + 4 = 14.
  - For [4,5][9,10], SDCM_ALL = (4-4.5)^2 + (5-4.5)^2 + (9-9.5)^2 + (10-9.5)^2 = 0.25 + 0.25 + 0.25 + 0.25 = 1.
  - For [4,5,9][10], SDCM_ALL = (4-6)^2 + (5-6)^2 + (9-6)^2 + (10-10)^2 = 4 + 1 + 9 + 0 = 14.
  - [4,5][9,10] has the smallest SDCM_ALL, so is "best ranges", minimizes variation within classes.

- **Step 3 – Cal**culate a "goodness of variance fit" (GVF), defined as (SDAM - SDCM) / SDAM. GVF ranges from 1 (perfect fit) to 0 (awful fit). Higher SDCM_ALL (more variation within classes) results in lower GVF.
  - In the examples in step #2, GVF is (26 - 1) / 26 = 25 / 26 = 0.96 for the best combination, and (26 - 14) / 26 = 12 / 26 = 0.46 for the two rejected combinations, a huge difference.

NATURAL BREAKS & JENKS OPTIMIZATION

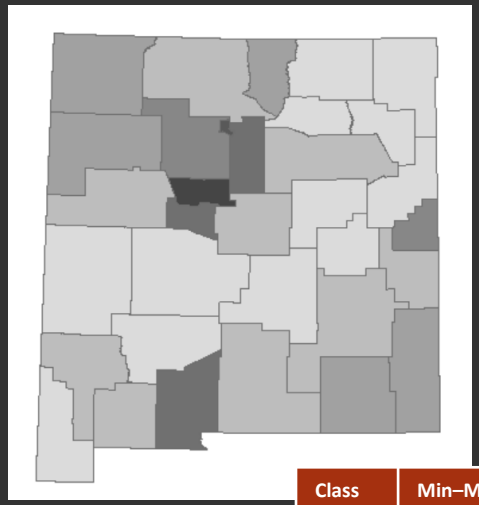
Ex: Population Density of 33 New Mexico Counties – 2010 Census

Natural Breaks



Class	Min–Max	Frequency
1	164.33 – 567.69	1
2	75.48 – 164.32	1
3	54.86 – 75.47	2
4	35.43 – 54.85	1
5	23.46 – 35.42	2
6	14.97 – 23.45	1
7	0.33 – 14.96	25

Natural Breaks (Jenks)



Class	Min–Max	Frequency
1	164.33 – 567.69	1
2	75.48 – 164.32	1
3	35.43 – 75.47	3
4	23.46 – 35.42	2
5	10.82 – 23.45	5
6	4.26 – 10.81	9
7	0.33 – 4.25	12

JENKS OPTIMIZATION

- Advantages
  - Attempts to minimize within class variance and maximize between class variance
  - High accuracy classification
- Disadvantages
  - Complex process
  - Difficult to understand procedure of grouping

NESTED MEANS

- Arithmetic mean of data is calculated to group the data into classes
- One above mean and one below mean
- Secondary means are calculated to divide those classes into two more classes

NESTED MEANS

- Advantages
  - Easily calculated
  - Mathematically intuitive
- Disadvantages
  - Limited to even no. of classes
  - Do not consider data distribution
  - Not included in GIS software

MEAN AND STANDARD DEVIATION

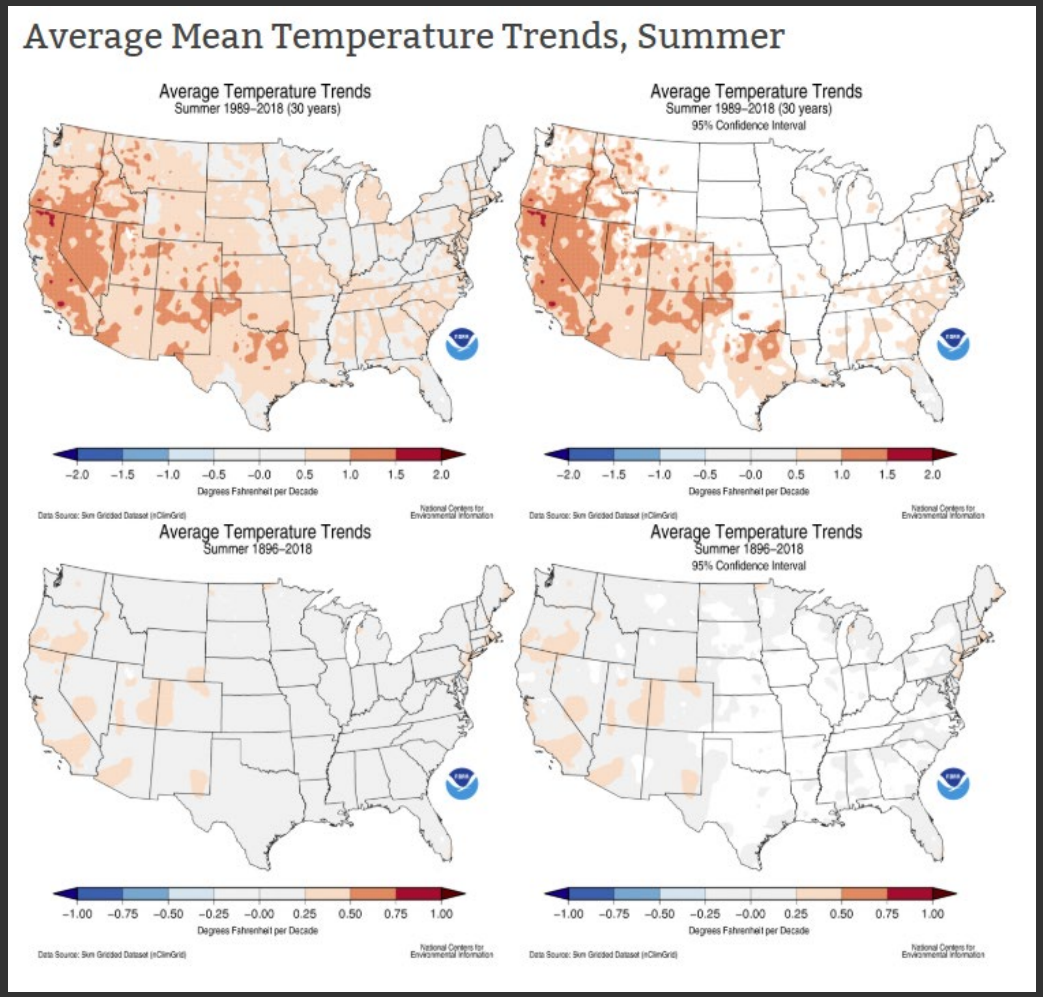
- Class boundaries are compiled by comparing mean and standard deviation, then determine boundaries by adding and subtracting the deviation from mean.
- Usually no more than 6 classes are needed

Ex: Population Density of New Mexico Counties – 2010 Census

Class	Min–Max	Frequency	Std. Dev.
1	0.32 – 85.31	31	< 0.5
2	85.32 – 184.63	1	0.5 – 1.5
3	184.64 – 567.68	1	> 1.5

Mean = 35.64
SD = 99.32

STANDARD DEVIATION



https://www.ncdc.noaa.gov/temp-and-precip/us-trends/tavg/sum

MEAN AND STANDARD DEVIATION

- Advantages
  - Good for data with normal distribution
  - Produces constant class intervals
  - Considers data distribution
- Disadvantages
  - Normal distribution is not common
  - Requires basic understanding of statistics
  - Not easily understood by the map reader
- Standard deviation is good for –
  - Visualizing features above or below an average
  - Displaying data that has a normal distribution

EQUAL INTERVAL

- Equal Step classification
- Data range of each class held constant
- Frequency of observations will vary between classes
- A zero frequency class should be avoided
- Outliers should be identified to avoid zero

EQUAL INTERVAL

Example: Population Density of 33 New Mexico Counties – 2010 Census

- Data Range = 567.36
- Number of Classes = 5
- Class Range = 113.47

Class	Min–Max	Frequency
1	454.23 - 567.69	1
2	340.75 - 454.22	0
3	227.28 - 340.74	0
4	113.81 - 227.27	1
5	0.33 - 113.80	31

EQUAL INTERVAL

- Advantages
  - Straight forward
  - Simple to compute
  - No gaps in the legend
- Disadvantages
  - Do not consider data distribution
  - May produce classes with zero observations
- Equal Interval maps are good for
  - Mapping continuous data. Ex: precipitation, temperature
  - Data with a fairly even distribution

EQUAL FREQUENCY OR QUANTILES

- Equal number of observations between classes
- Quartiles (4 classes) or Quintiles (5 classes)
- Example: Population Density of 33 New Mexico Counties – 2010 Census

Number of Classes = 5
Observations/class = 6.6

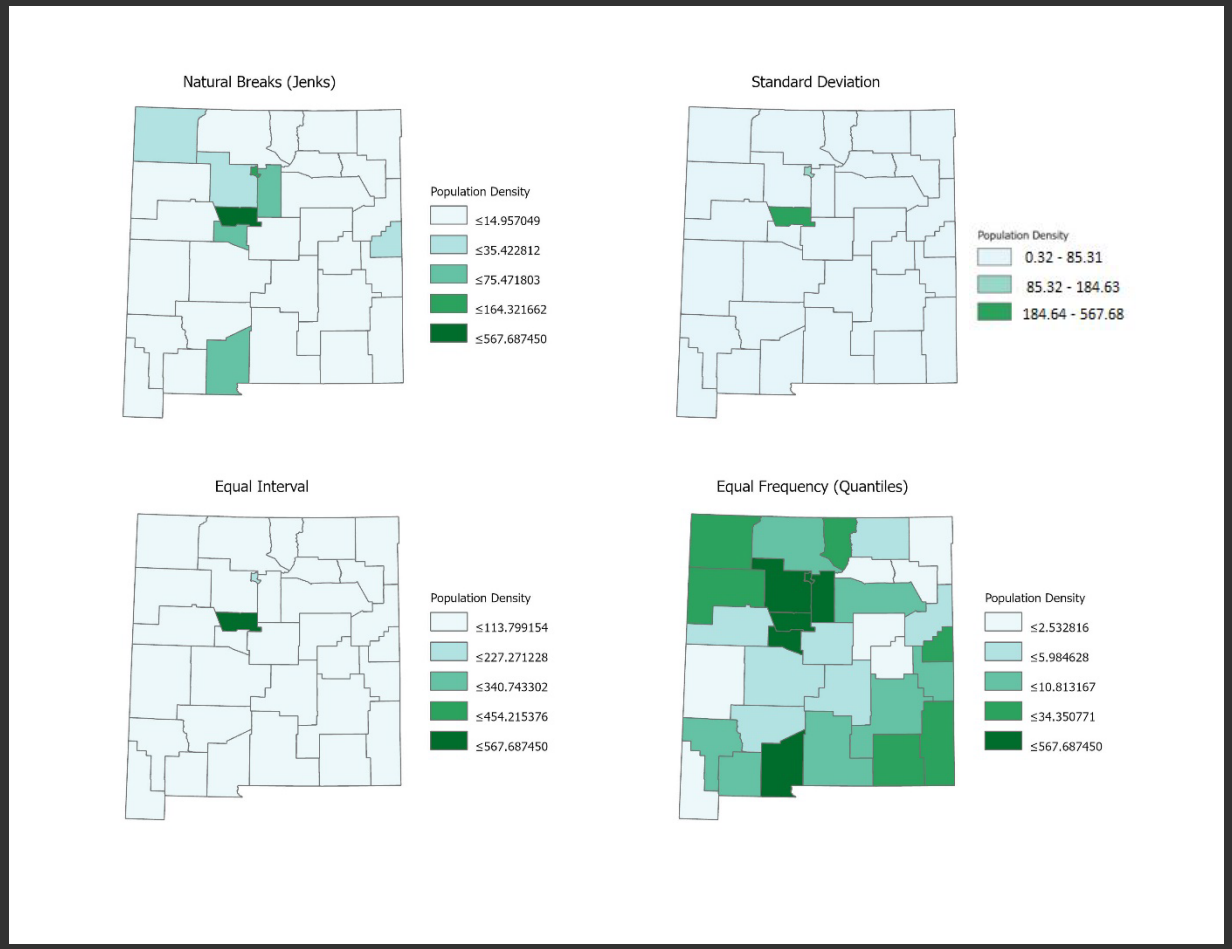
Class	Min –Max	Frequency
1	34.36 - 567.69	6
2	10.82 - 34.35	6
3	5.99 - 10.81	7
4	2.54 - 5.98	7
5	0.33 - 2.53	7

EQUAL FREQUENCY OR QUANTILES

- Advantages
  - Easily calculated
  - No empty classes
- Disadvantages
  - Does not consider data distribution
  - Value of observation could be close to values in a different class
  - Distribution is not equal if no. of classes is not a whole number.
- Quantiles is good for
  - Data with a fairly even distribution
  - Emphasizing the relative positions of a feature among other features; ex: quantile will clearly show which counties are in the lowest or highest 20%

COMPARING VARIOUS METHODS

Ex: Population Density of 33 New Mexico Counties – 2010 Census



ARITHMETIC AND GEOMETRIC INTERVALS

- Arithmetic
  - a, a+d, a+2d, a+3d,…, a+(n-1)d
  - a = first term, d = common difference, n = no. of terms
- Geometric
  - a, ar, a(r^2), a(r^3),…, a(r^(n-1))
  - a = first term, r = common ratio

ARITHMETIC AND GEOMETRIC INTERVALS

- Advantages
  - Good for data with large ranges
  - Break points determined by rate of change in data
- Disadvantages
  - Not appropriate for data with small ranges or linear trends

USER DEFINED

- Not a true classification system
- Advantages
  - Flexibility for the user
  - Break points may be defined based on the spatial distribution
- Disadvantages
  - Logic of legend breaks not apparent
  - Not easily repeated