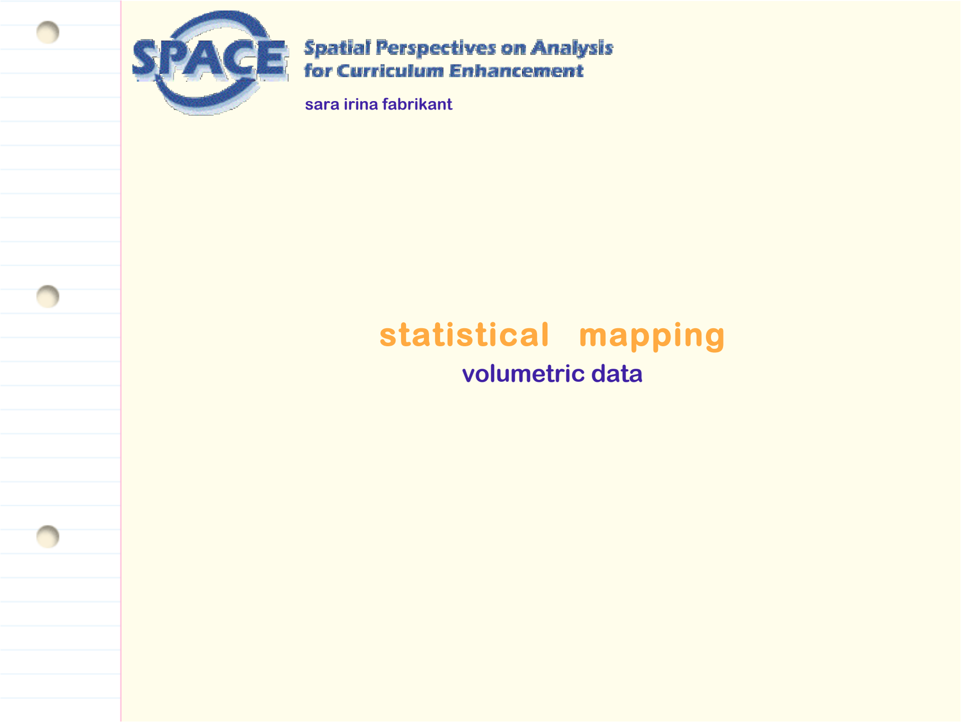


Spatial Perspectives on Analysis for Curriculum Enhancement

sara irina fabrikant

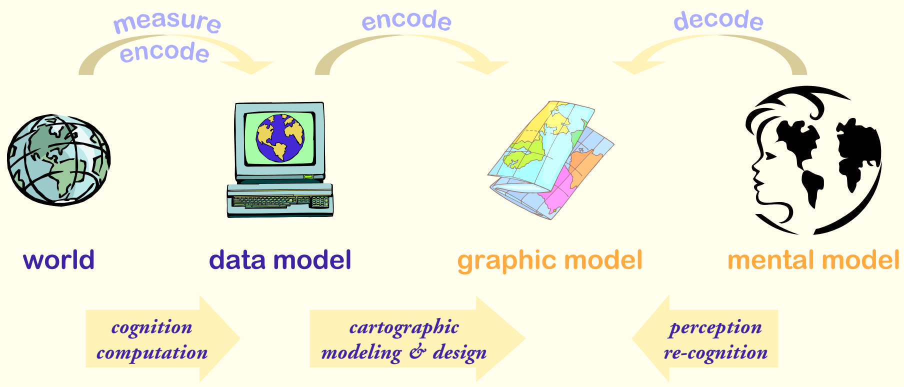
statistical mapping

volumetric data

outline

- volumetric data
  - areas: choropleth
- classification
  - to class or not to class?
  - evaluate classification solution
- design issues
  - legend
  - color

cartographic process



measure
encode

encode

decode

world

data model

graphic model

mental model

cognition
computation

cartographic
modeling & design

perception
re-cognition

data model vs. graphic model

goal: graphic model vs. data model ➔best fit

- data model: volumes (continuous, 3D) at points & areas

2D gm

3D gm

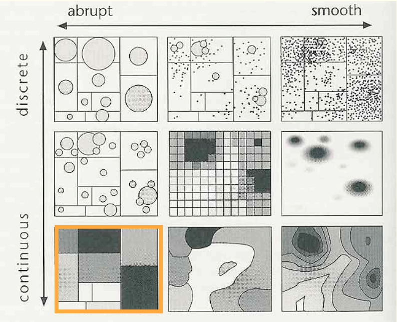


FIGURE 6.53. A syntactics of map forms. *Reproduced from* MacEachren (1992b, *Fig.* 9, *p.* 16). *Reprinted by permission of* Cartographic Perspectives.

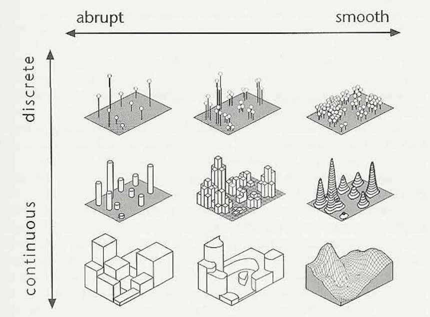
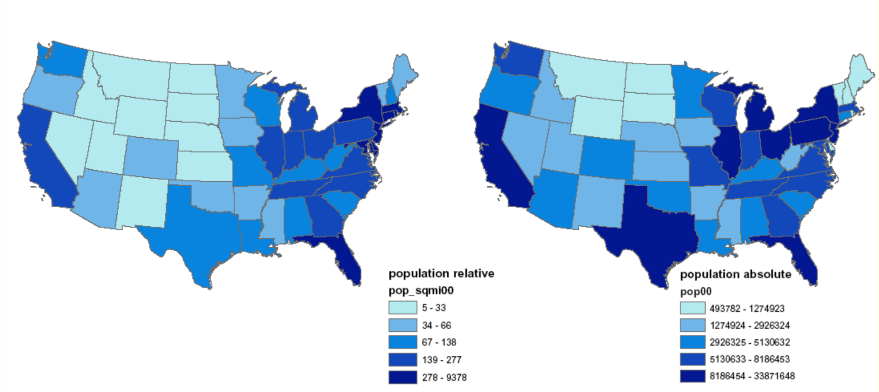


FIGURE 6.52. A typology of data models. *Reproduced from* MacEachren (1992b, *Fig.* 8, *p.* 16). *Reprinted by permission of* Cartographic Perspectives.

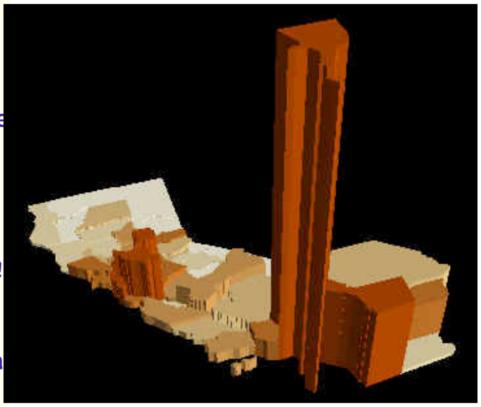
MacEachren (1995)

U.S. population in 2000 (volumes)



volumetric data in areas

- choropleth map
  - choros = place, space
  - plethos = magnitude
- continuous data: ratios, densitie
- discrete graphic model
  - stepped surface
  - boundaries unrelated to data
  - adjust data model: standardize!
- good for…
  - finding value of a given area
  - gist of overall pattern
  - compare patterns between n



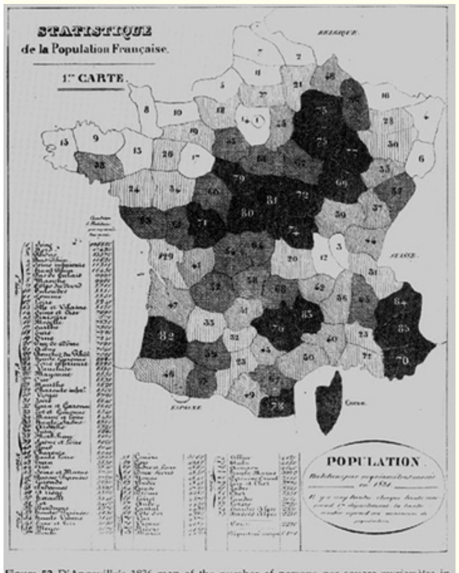
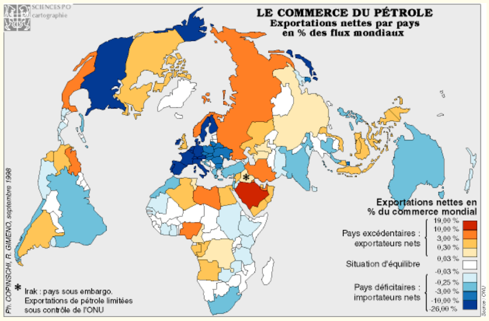
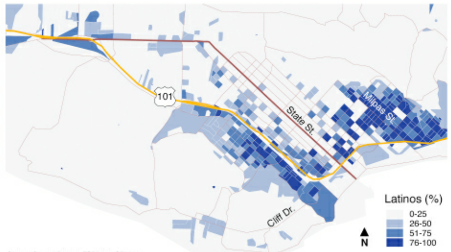


Figure 52 D'Angeville's 1836 map of the number of persons per square *myriamètre* in France. Original 187 × 239 mm. Lithograph. (Photo. Bibliothèque Nationale, Paris.)

source: D'Angeville, 1836, in: Robinson



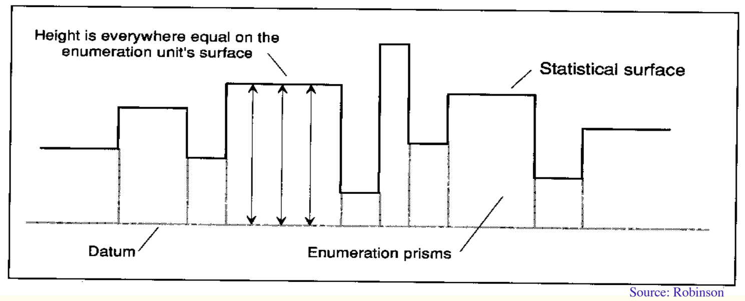
Proportion of Latinos in Santa Barbara



Source: Census Bureau, ESRI data CD 2000

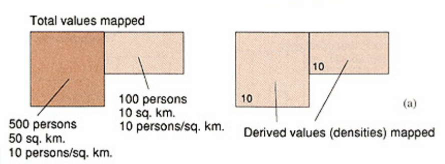
Authors: Sara Fabrikant, Stuart Sweeney and Steven Hochart of the UCSB Geography Department

discrete statistical surface model



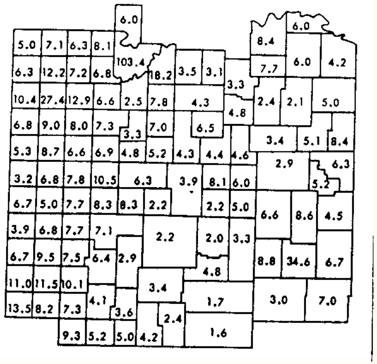
Source: Robinson

- do NOT use raw data
- but a variable per unit per area



Source: Dent

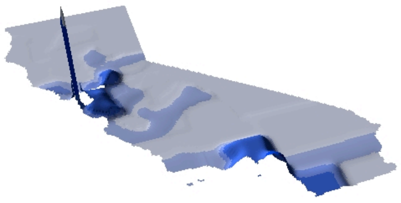
raw map data ➔ 3D symbolization space



Source: Robinson

graphic representations of map data

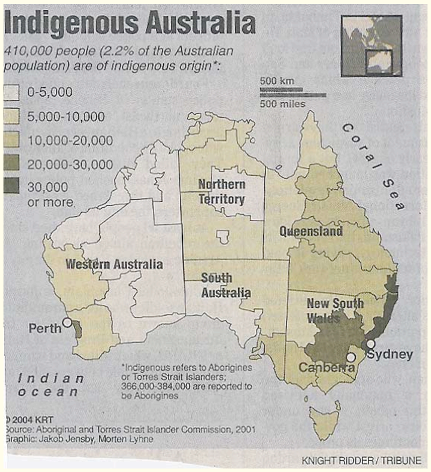
- 3D representations
- discrete surface maps
  - pin head maps
  - 3D bar maps
  - prism maps
- continuous surface maps
  - e.g using centroid of unit area
    - various interpolation methods
  - e.g. using boundary unit area
    - pycnophylactic interpolation



sara fabrikant, 2001

outline

- volumetric data
  - areas: choropleth
- classification
  - to class or not to class
  - evaluate classification solution
- design issues
  - legend
  - color



Indigenous Australia

410,000 people (2.2% of the Australian population) are of indigenous origin:*

*Indigenous refers to Aborigines or Torres Strait Islanders; 366,000-384,000 are reported to be Aborigines

© 2004 KRT
Source: Aboriginal and Torres Strait Islander Commission, 2001
Graphic: Jakob Jensby, Morten Lyhne

KNIGHT RIDDER / TRIBUNE

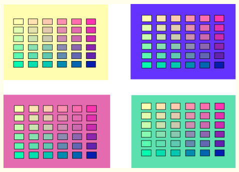
choropleth map types

classed choropleth maps

- data more aggregated (stat. error ↑)
- perceptual limits of how many categories can be perceived
  - not more than 11 area-shaded gray tones
- 5-7 classes appropriate most of the time (perc. error ↓)
  - (Miller: 7 ± 2)
- if animated, closer to 3 classes

unclassed choropleth maps

- data less aggregated (stat. error ↓)
- many individual values (perc. error ↑)
  (based on empirical findings)



unclassed choropleth map

pro: (e.g., Tobler)

- stat. data = continuous
- data model = graphic model

con (e.g., Dobson)

- graphic model <> mental model
- map <> legend
- distribution dependent
- map comparison is hard

solution:

- ESDA: make both types!

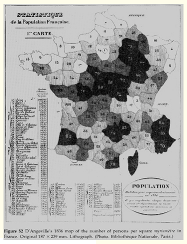
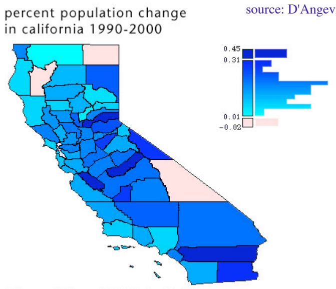


Figure 52 D'Angeville's 1836 map of the number of persons per square *myriamètre* in France. Original 187 × 239 mm. Lithograph. (Photo. Bibliothèque Nationale, Paris.)

source: D'Angeville, 1836, in: Robinson



data source: US Census, 2001, ESRI shapefiles 2000.

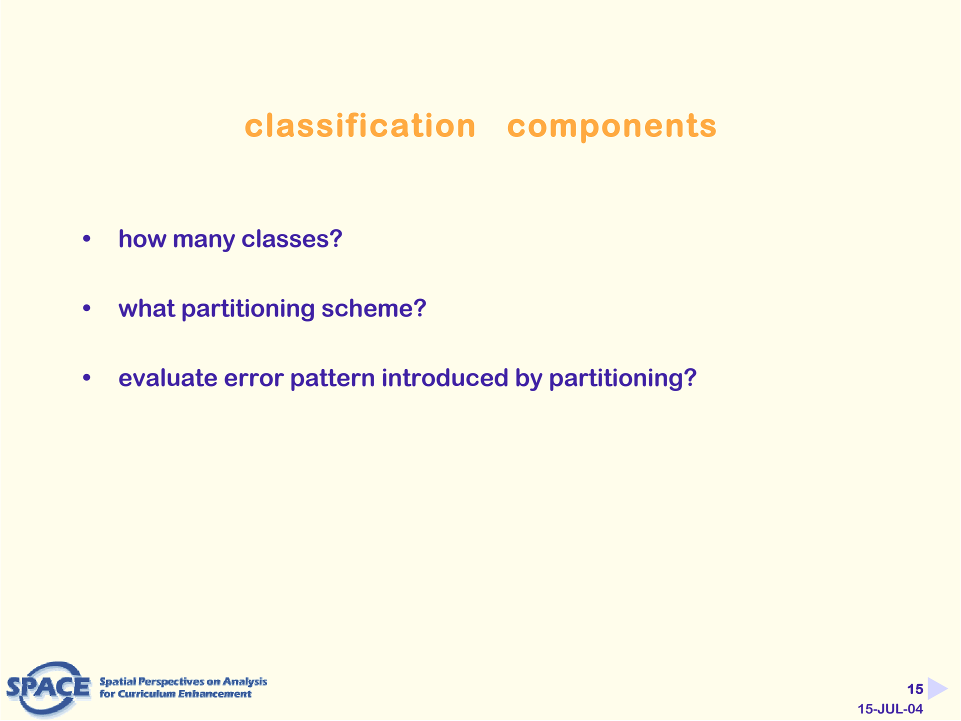
sara fabrikant, 2001

to class or not to class?

classing most useful if distribution (Evans, 1977)...

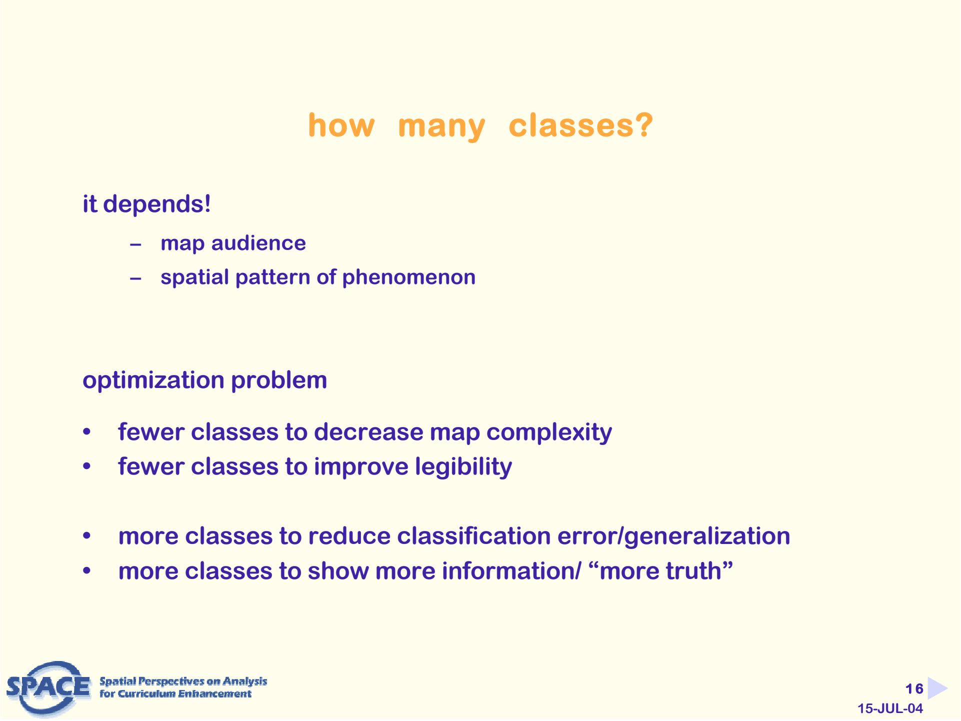
- shows natural breaks
- is multimodal
- is in some progression
- of phenomena show concrete breaks or distinctions
  - (e.g., people, buildings etc.)

- classing is useful because this is how the brain works
  - categorization



classification components

- how many classes?
- what partitioning scheme?
- evaluate error pattern introduced by partitioning?



how many classes?

it depends!

- map audience
- spatial pattern of phenomenon

optimization problem

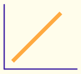
- fewer classes to decrease map complexity
- fewer classes to improve legibility

- more classes to reduce classification error/generalization
- more classes to show more information/ “more truth”

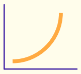
what partitioning scheme?

guess what: it depends!

- idiographic classes based on the nature of the data?
  - e.g. look for natural patterns in data
- arbitrary classes with round numbers?
  - e.g. 10-20, 21-40, etc.



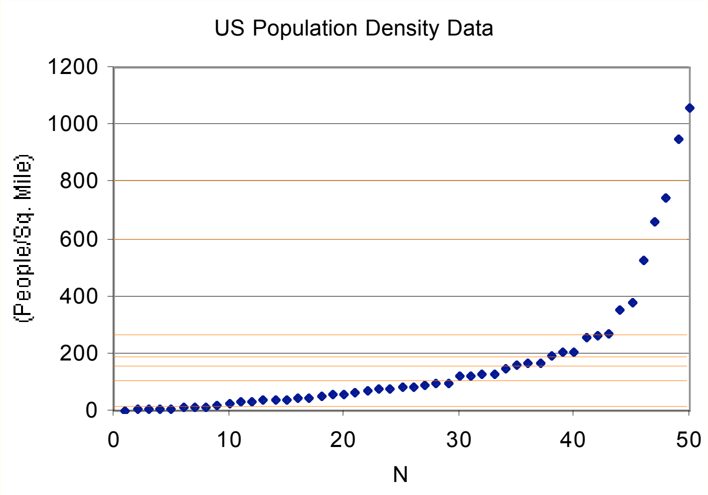
- serial classes based on mathematical principles?
  - e.g. geometric progression



- exogenous classes based on a related variable?
  - e.g. income based on a “poverty” variable

recipe…

- start by creating a graph of your data (lab: part I !)
- rank your data: plot lowest to highest values



ideographic schemes

Greek - “descriptor of uniqueness”

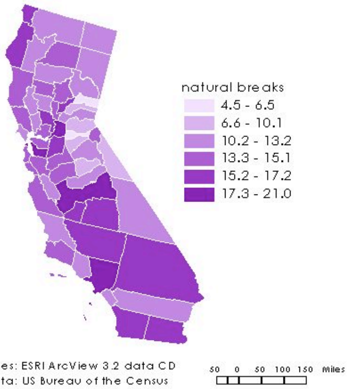
- clinographic - natural breaks
  - look for discontinuities in array (data unevenly distributed)
- quantiles based (n-tiles)
  - data values evenly segmented (data evenly distributed, compare ranks)
- contiguous
  - spatially homogeneous (data spatially correlated)
- correlation based
  - high similarity (data semantically correlated)

SPACE Spatial Perspectives on Analysis for Curriculum Enhancement

natural breaks

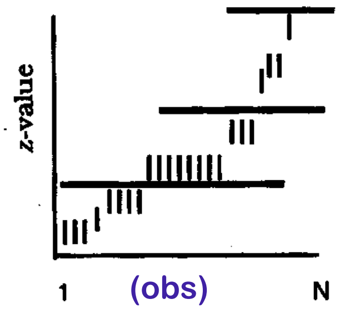
example...

female headed households in california counties in 1999

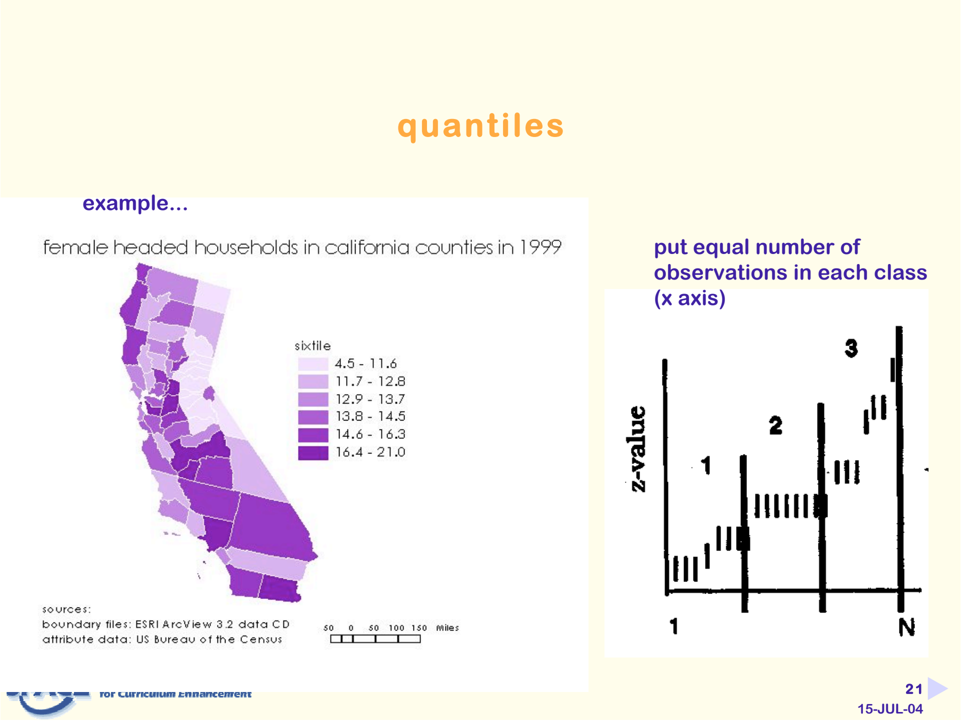


sources:
boundary files: ESRI ArcView 3.2 data CD
attribute data: US Bureau of the Census

look for gaps in
the array of values (y-axis)



for Curriculum Enhancement



quantiles

example...

female headed households in california counties in 1999

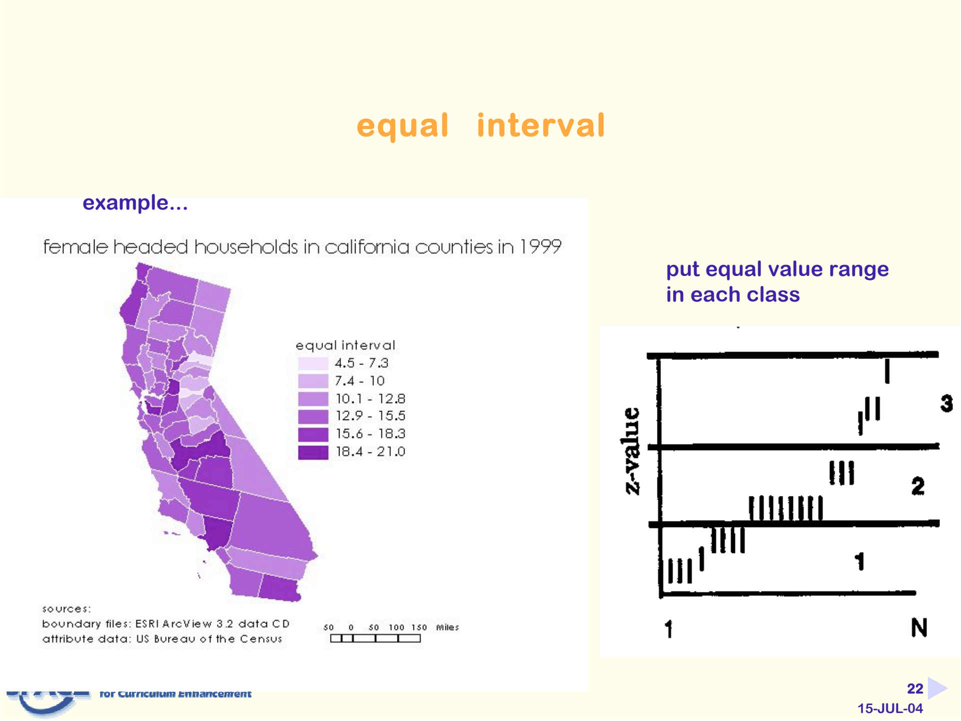
sixtile
- 4.5 - 11.6
- 11.7 - 12.8
- 12.9 - 13.7
- 13.8 - 14.5
- 14.6 - 16.3
- 16.4 - 21.0

sources:
boundary files: ESRI ArcView 3.2 data CD
attribute data: US Bureau of the Census

put equal number of observations in each class (x axis)

for Curriculum Enhancement

21
15-JUL-04



equal interval

example...

female headed households in california counties in 1999

equal interval
- 4.5 - 7.3
- 7.4 - 10
- 10.1 - 12.8
- 12.9 - 15.5
- 15.6 - 18.3
- 18.4 - 21.0

sources:
boundary files: ESRI ArcView 3.2 data CD
attribute data: US Bureau of the Census

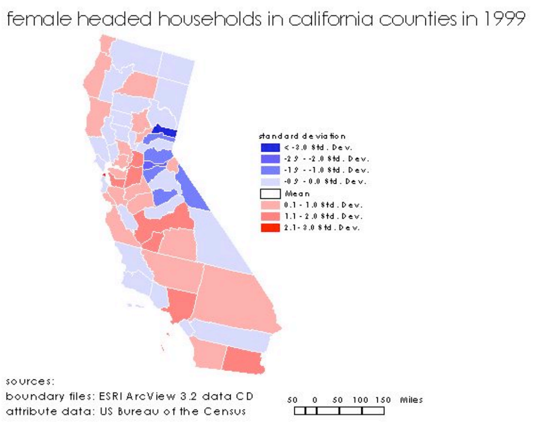
put equal value range in each class

for Curriculum Enhancement

22
15-JUL-04

standard deviation

example...



female headed households in california counties in 1999

sources:
boundary files: ESRI ArcView 3.2 data CD
attribute data: US Bureau of the Census

average of the absolute deviations from the mean

all distributions
m +/- 2s = 75% obs.
m +/- 3s = 80% obs.

normal distribution
m +/- 1s = 68% obs.
m +/- 2s = 95% obs.
m +/- 3s = 99% obs.

optimization method

the problem:

- for 10 values (n) and 3 groups (r) there are 36 ways to contiguously partition the data!

$$\frac{(n-1)!}{(r-1)!\left[\left(n-1\right)-(r-1)\right]!}$$

goal:

find exhaustive set of distinctly different groups, while keeping groups internally most homogeneous

➔ optimization problem (e.g., cluster analysis)!

what class breaks?

class breaks such that...

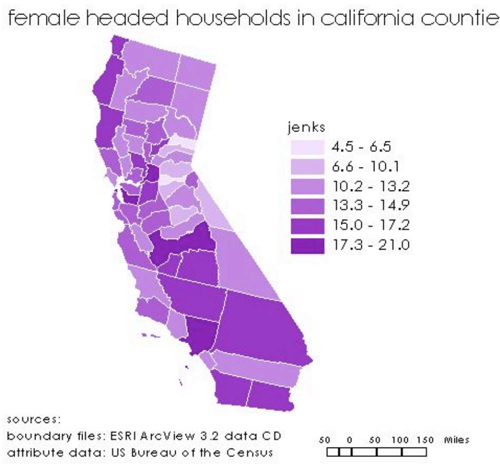
- similarity within the classes is maximal, &
- differences between the class is maximal.
  - natural breaks (–> where large gaps occur)
- iterative mathematical procedure
  - find a global minimum of within-class dispersion, and find a global maximum of between-class dispersion
- least squared distance from the (class) mean
  - minimize the blanket of error:
  - smallest sum of weighted squared deviations from the class mean

$$\text{Dist.}_{\min} = w_i \sum_{j=1}^{N} \left(z_i - \bar{z}_i \right)^2$$

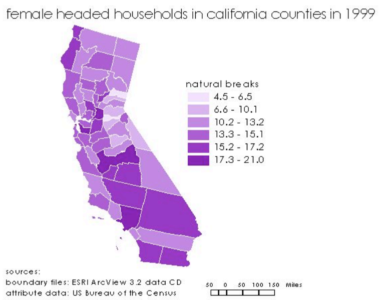
area obs$_i$

SPACE Spatial Perspectives on Analysis for Curriculum Enhancement

optimized solution


female headed households in california counties in 1999

sources:
boundary files: ESRI ArcView 3.2 data CD
attribute data: US Bureau of the Census


female headed households in california counties in 1999

sources:
boundary files: ESRI ArcView 3.2 data CD
attribute data: US Bureau of the Census

SPACE for Curriculum Enhancement

goodness of variance fit - GVF

Jenks (1977)

- second step to maximize distance between classes

GVF = sum of squared deviations between classes

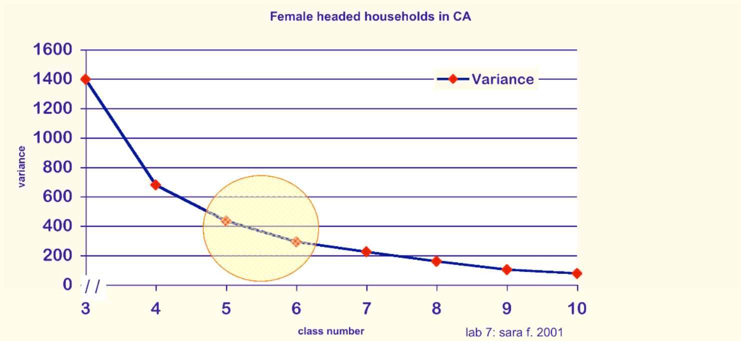
- compute squared deviations from overall mean (data array) and subtract squared deviations from the class means

- go trough various iterations until GVF is maximized

- Range of GVF
  - max. 1.0 = each value is one class –> unclassed choropleth map
  - min. 0.0 = one class only
  - ideal: closest to one as possible

how many classes?

- minimize number of classes without loosing too much information
- find elbow in the variance graph…



lab 7: sara f. 2001

evaluation of partitioning method

Jenks and Coulson (1963)

- visual check on partitioning validity
  - remember choropleth data model: a statistical surface
- compute discrepancy between each value and its associated class mean

Blanket of Error

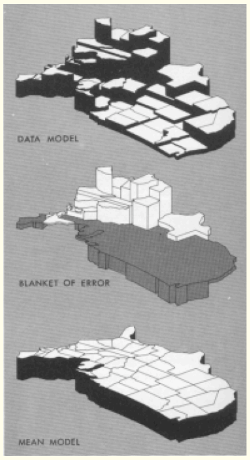
- the classification error = statistical surface (the error values fluctuating above and below the class mean)
- akin to root mean square error (RMSE) e.g., standard deviation from mean

$$\mathrm{RMSE} = \sqrt{\frac{\sum\left(x_i - \bar{X}\right)^2}{N}}$$

blanket of error

Jenks (1967):

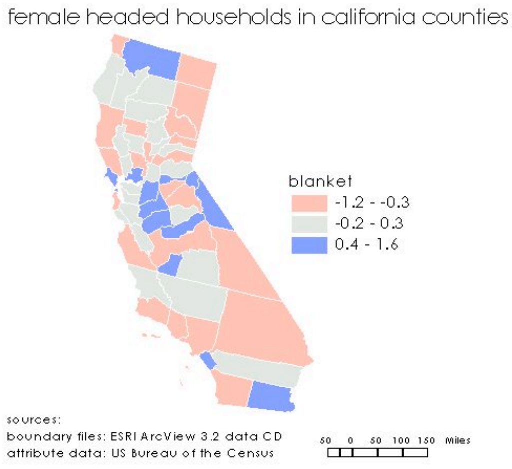
> *"We have found in our study, however, that the series of classes with minimal error and those with an uniformly distributed error are not significantly different statistically.*
> *As a result we assume that the cartographer should use equal average or equal relative deviation classes for choroplethic maps."*



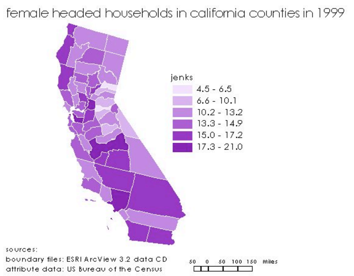
evaluation (cont.)

- what is the error pattern on the map?
  - is the blanket of error uniformly distributed?
  - is concentrated in certain areas and not in others?
  - if yes, why?
- error measure is sensitive to # of classes and classing scheme!
- with optimized classing, the error should be minimized
  ➔ Natural Break (Jenk's) method in ArcMap is minimizing
- cartographer controls this error by modifying classing scheme!

blanket of error (cont.)


female headed households in california counties in 1999

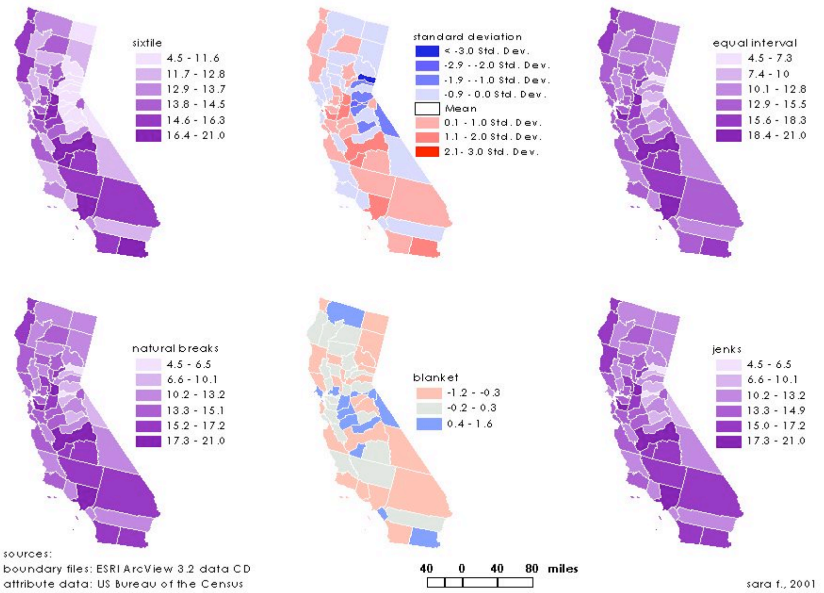
sources:
boundary files: ESRI ArcView 3.2 data CD
attribute data: US Bureau of the Census


female headed households in california counties in 1999

sources:
boundary files: ESRI ArcView 3.2 data CD
attribute data: US Bureau of the Census

female headed households in california counties in 1999

identical data, same pattern?



sources:
boundary files: ESRI ArcView 3.2 data CD
attribute data: US Bureau of the Census

sara f., 2001

classed choropleth maps

- summary: many possible schemes !

which solution is best?

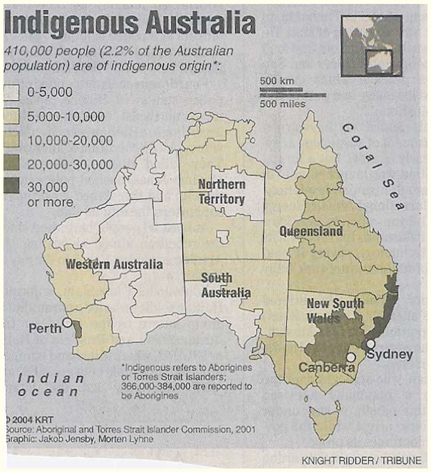
- it depends on the data (always inspect distribution!)
- it depends on the scale (ecological fallacy, MAUP)
- several good solutions possible!

data requirements

- based on enumeration unit (e.g. census tract)
- no totals, as enumeration units vary in size!
- standardized data (ratios of some sort)
  - density (area dependent)
  - per capita income (area independent)

outline

- volumetric data
  - areas: choropleth
- classification
  - to class or not to class
  - evaluate classification solution
- design issues
  - legend
  - color



*Indigenous refers to Aborigines or Torres Strait Islanders; 366,000-384,000 are reported to be Aborigines

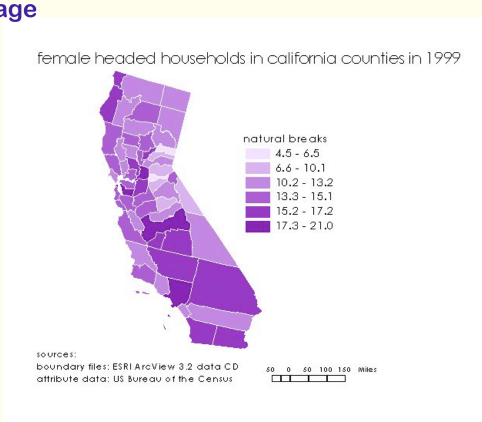
© 2004 KRT
Source: Aboriginal and Torres Strait Islander Commission, 2001
Graphic: Jakob Jensby, Morten Lyhne

KNIGHT RIDDER / TRIBUNE

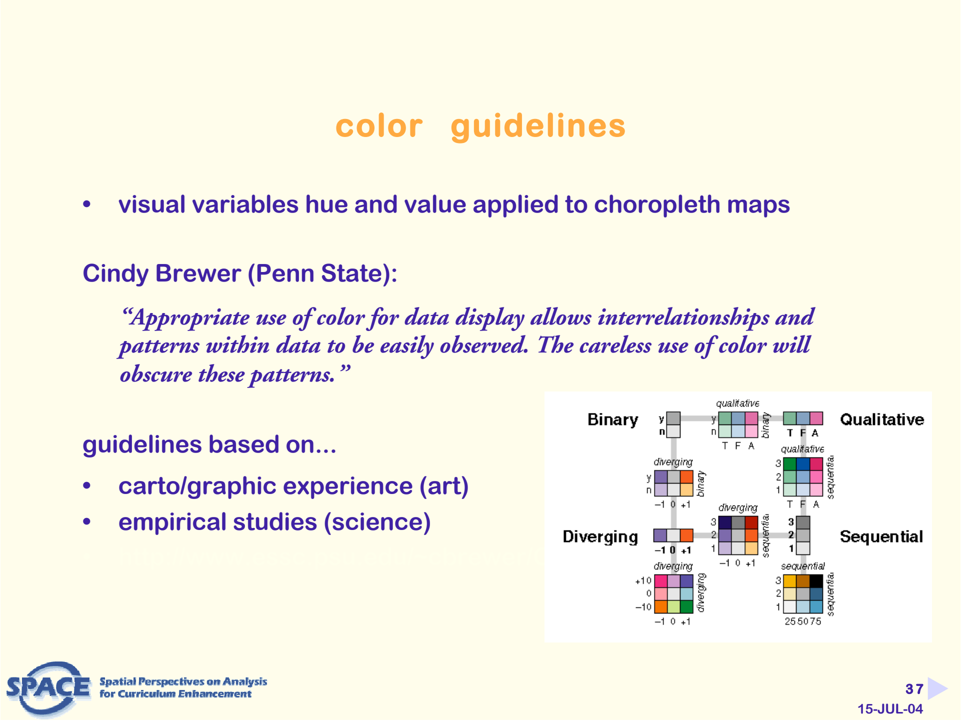
legend design: an important side note…

- map and legend on the same page
- correct labeling of classes
  - simple, concise and legible
- beware of software ‘defaults’
  - e.g. fix ESRI-isms
- all values have to be classed
  - no gaps, no overlapping bins
- example…

  0 – 10 can be labeled: < 11
  11 – 20
  21 – 40
  41 – 60 can be labeled: > 40



sources:
boundary files: ESRI ArcView 3.2 data CD
attribute data: US Bureau of the Census



color guidelines

- visual variables hue and value applied to choropleth maps

Cindy Brewer (Penn State):

> *"Appropriate use of color for data display allows interrelationships and patterns within data to be easily observed. The careless use of color will obscure these patterns."*

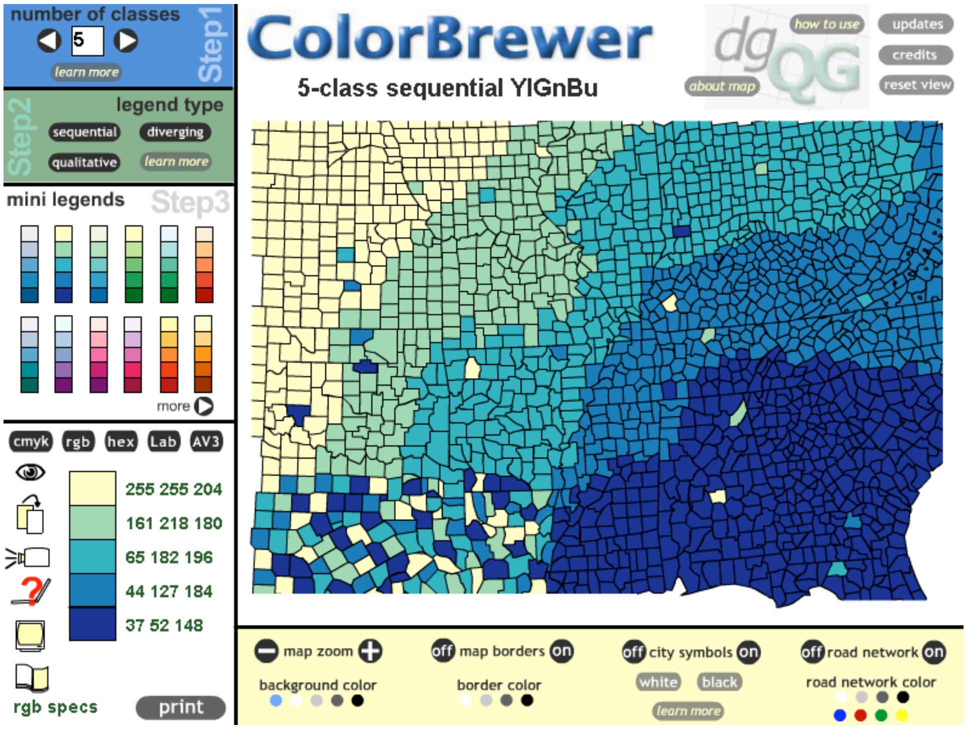
guidelines based on...

- carto/graphic experience (art)
- empirical studies (science)

Spatial Perspectives on Analysis for Curriculum Enhancement

37

15-JUL-04



ColorBrewer

5-class sequential YlGnBu

255 255 204
161 218 180
65 182 196
44 127 184
37 52 148