L4: Thematic mapping

Kraak & Ormeling, Cartography – Visualization of Geospatial Data
- chapter 7: Statistical mapping

Statistical data collection

Statistical surveys of socio-economic data

Collection of socio-economic data

- continuously
- Census
  - population (10 years)
  - health (good surveys developed)
  - industrial activity
  - education and schooling
- Sampling frameworks (discrete)

Privacy regulations: data are combined before publishing.
Information on individual households, farms, plants, companies can not be worked out from the data.

Statistical data collection

Statistical surveys of physical, chemical and biological data

Collection of physical, chemical and biological data

→ continuously (RS)

→ Sampling programs
  - → Water flow and chemistry
  - → Soil conditions
  - → Biomass production (carbon mass balance)
  - → Agricultural production

→ Sampling frameworks (discrete)

Privacy regulations: data are combined before publishing.
Information on individual farms and plots can not be worked out from the data (dependent on legal regulations in country of study).

Data analysis

After the collection of the data, these have to be analysed in order to choose the correct method for their representation and visualisation.

1. Assessing validity of the data:

- when were the data collected?
- in which way?
- for what purpose?
- for which period of time?
- to what area do they refer?
- are they comparable to older data (in order to realize a time series)?

} Usefulness, reliability and accuracy of data

The description of the data is called metadata, and ideally the metadata should always follow the data, e.g. in the data-file header or as a separate file.

Data analysis

2. Assessing data characteristics:

- the nature of objects the data refer to (point, linear, areal, volumetric objects)

- the type of change in the data (gradual, abrupt), related to distribution

distribution →
- continuous (measured everywhere - air pressure, temperature)
- discrete (only at certain locations – land cover classes, vegetation classes, bedrock classes)

- the measurement scale (nominal, ordinal, interval, ratio)

Data analysis

Data measurement scales

- Nominal scale → Differences in data qualitative → gender, language, religion, land use, geology
- Ordinal scale → Differences in data hierarchical, there is a natural order → cold-moderate-warm-hot, small-medium-large
- Interval scale → Order + distance → Temperature: ºC or K scale, Vegetation index
- Ratio scale → Order + distance + comparison between values → Population density: no. of inhabitants/km^2

Data analysis

Ratio scale

Absolute ratio: result of direct measurements or addition of units

income (in SEK) – continuous variable
no. of children in a family – discrete variable

Relative ratio: absolute ratio data related to other data sets

putting the absolute ratio data into context

- index values
- averages
- ratios
- densities
- potentials

Data analysis

Index values for time series:
- how much in today's money was 10000 SEK in 1950?

Density:
- ratio between the population or an area and the resources available to that population (either the residential area or the agricultural area they are cultivating)

Non-area-related ratios:
- relationship between any two data sets (no.of doctors/population),
- or relationship between two subsets of population (no.of doctors/no.of teachers)
- example:

total number of influenza patients in Sweden = 100000
total number of influenza patients in Italy = 120000

But, Sweden has ca. 9 million inhabitants and Italy ca. 60 million inhabitans!
A more objective comparison is to compare the ratios:

ratio of influenza patients/population =
ca. 1% (Sweden) = 0.2% (Italy)

Data analysis

Averages:

- characterisation of a data set by one number
- this is only successful for data with small variation in measurement
- three different average measures:

A series of given data values: 1, 1, 1, 1, 2, 3, 5, 10, 100

mean = sum of all values / number of all values = (1+1+1+1+2+3+5+10+100)/9 = 13.77

median = the middle value (50% of all values are larger/smaller than this value) = 2

mode = the most frequent value = 1

Data analysis

Nearest neighbour index:

- distribution patterns of point locations, the topological charateristics of line patterns, the shape of areal patterns
- R_n = comparison between random patterns and actual pattern:

R_n =
- 0 -> all observations in the actual pattern are in one point
- 1 -> actual pattern is a totally random one
- 2.15 -> actual pattern is completely regular (distances between all points equal)

Table 1 Nearest neighbour index values of places over 10 000 inhabitants per province in the Netherlands

Drenthe	1.6
Overijssel	1.5
Limburg	1.20
Friesland	1.18
Noord-Holland	1.16
Noord-Brabant	1.15
Gelderland	1.08
Zeeland	1.04
Zuid-Holland	1.01
Utrecht	1.0
Groningen	0.93
Flevoland	2.1

Figure 7.4 Population centres with over 10 000 inhabitants in the Netherlands

Data analysis

Potentials:

- potential in physics = attraction between two masses = $(m_1 * m_2) / d(m_1,m_2)$

- in geography:

  virtual interaction between

    the inhabitants of different cities = population potential

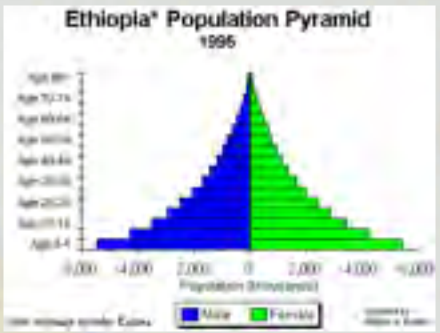
    expected purchases in a market = market potential

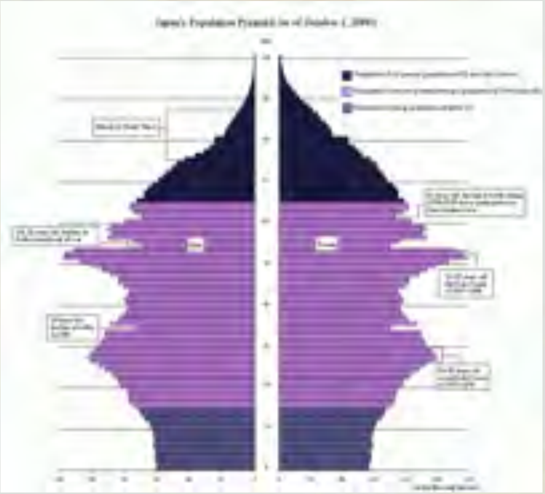
- population potential at a certain location = chance that the people at that location would meet people from other locations (neighbour cities)

Interval scale of human interactions in the UK and Ireland

Data adjustment

Normalisation – to minimize the distorting effects of irregularities in the population structure or geographical features



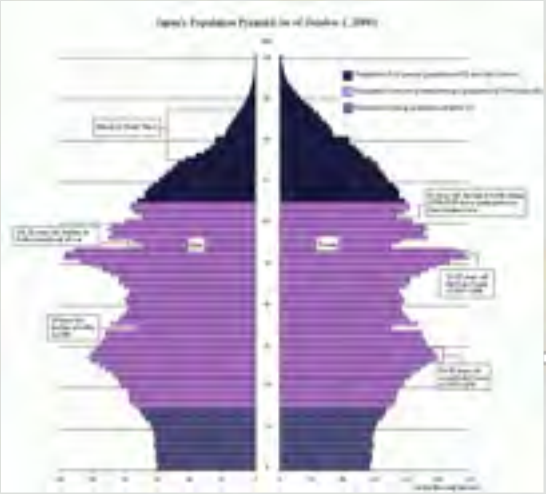


Data adjustment

Example 1:
- birth rate = number of births / 1000 inhabitants
- death rate = number of deaths / 1000 inhabitants

But birth and death rate depends on the form of the population pyramide.

In an area with a large amount of old people, death rate will be higher and birth rate lower than in an area with a normal population.



Data adjustment necessary!

- fertility rate = number of births / 1000 women in child-bearing age

Data adjustment

Example 2:
physical geography - minimizing effects of relief upon climate

0.6^{o}C degree decrease in temperature for each 100m of elevation

↓

All temperatures can be adjusted to their sea-level values.

Data classification

Mapping unprocessed data → Unclear visualisation

↓

Data classification: systematical grouping of data based on one or more characteristics

↓

Clearer map image

The data classification can be done either by producing a new thematic layer, or by symbolisation of the original data into discrete classes.

Data classification

Limit the number of classes: humans can handle approx. 7 classes to get an overview and and understanding the mapped theme at a glance.

Statistical data

Classification

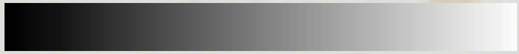
Legend: several grey shades

Advantages:
- a limited number of grey shades improves the legibility of the map

Disadvantages:
- data is generalised

No classification

Legend: continuous grey scale

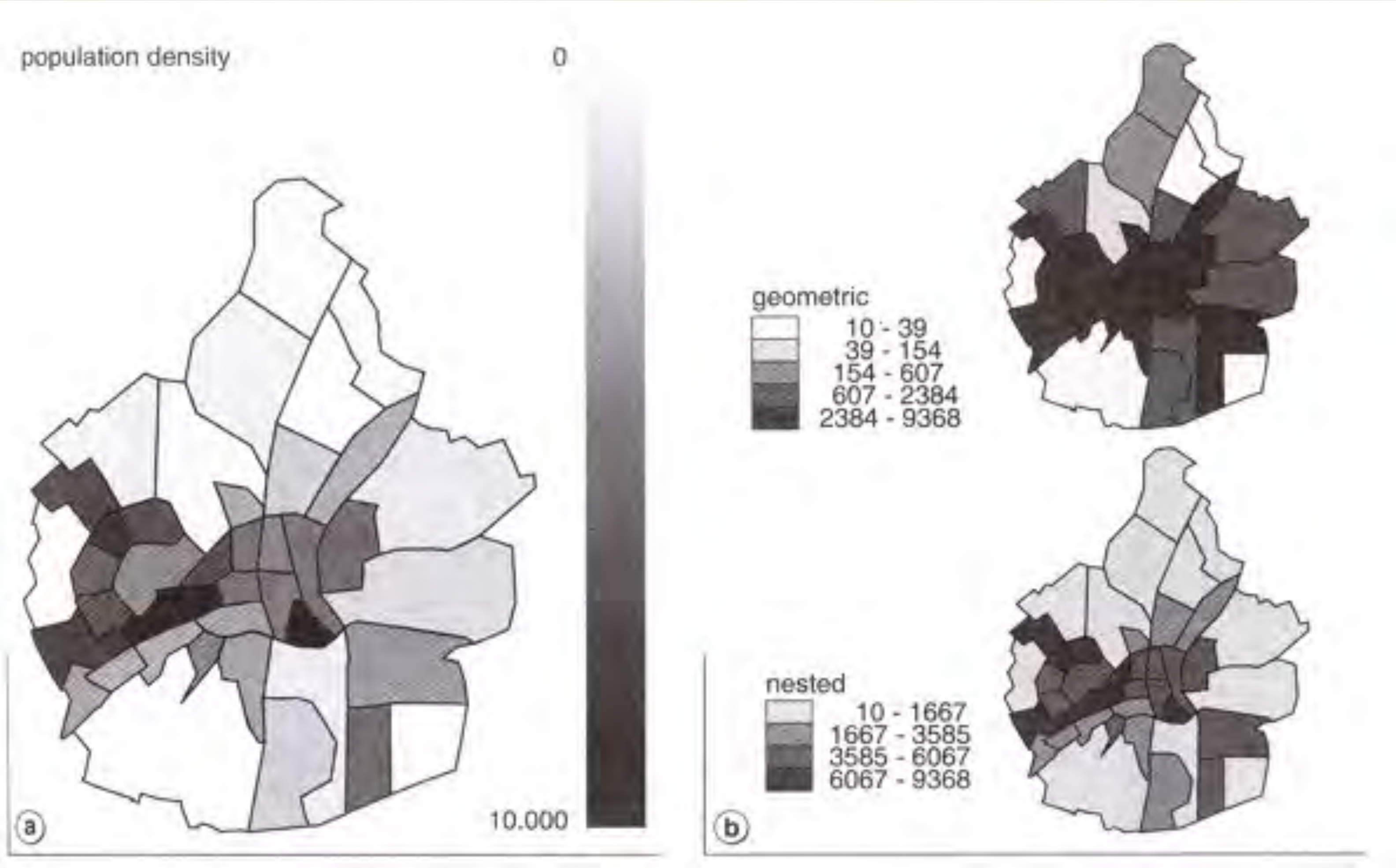


Advantages:
- the resulting image is not generalised

Disadvantages:
- difficult to perceive small differences if the areas are far away from each other
- not possible to do in most software packages

Data classification

Classification vs. no classification



Data classification

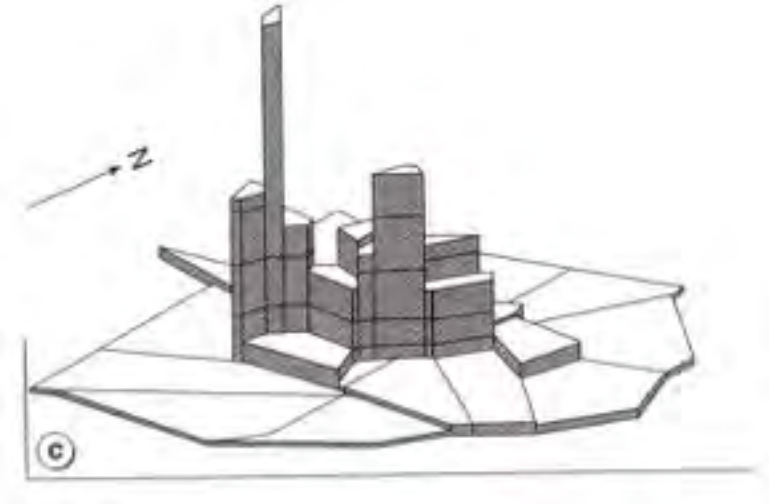
Classification requirements:

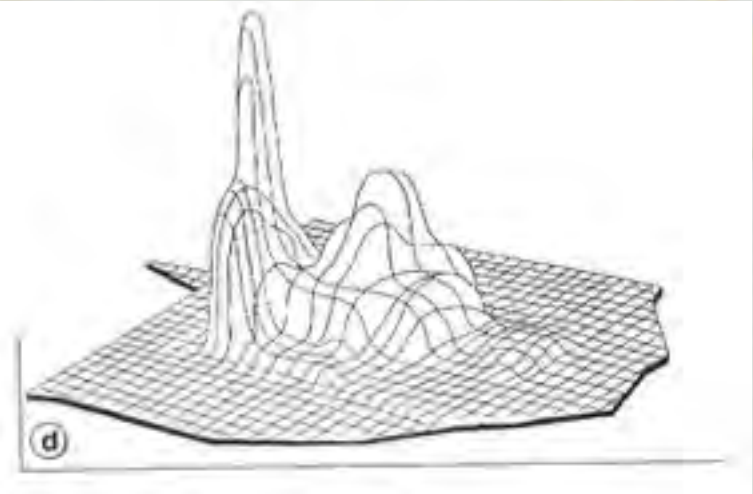
1. The final map has to be as close to the actual statistical surface as possible:

Statistical surface:
3D representation of the data, z = numerical value of the attribute

stepped surface
(choropleth map)

continuous surface
(isoline map)





Data classification

Classification requirements:

2. The final map should display the patterns/structures, which are characteristic for the displayed phenomena. Extreme values should not disappear through classification method.

3. Each class should contain observed values.

If these requirements are met:
- map gives a clear overview of the phenomenon,
- it is possible to determine value of the mapped attribute at every location on the map.

Data classification

Classification in 3 steps:

1. Choose a map type.

2. Limit the number of classes.

3. Define the class limits – the most difficult step.

Graphical methods

Break points
Frequency diagram
Cumulative frequency diagram

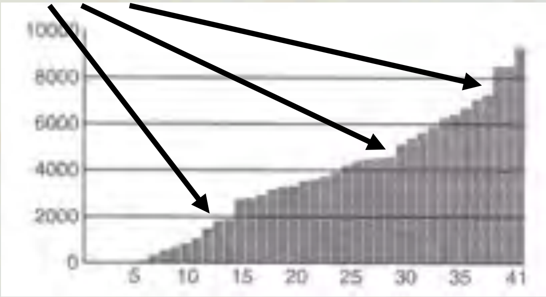
Mathematical methods

Equal steps
Quantiles
Arithmetic series
Geometric series
Harmonic series
Nested means

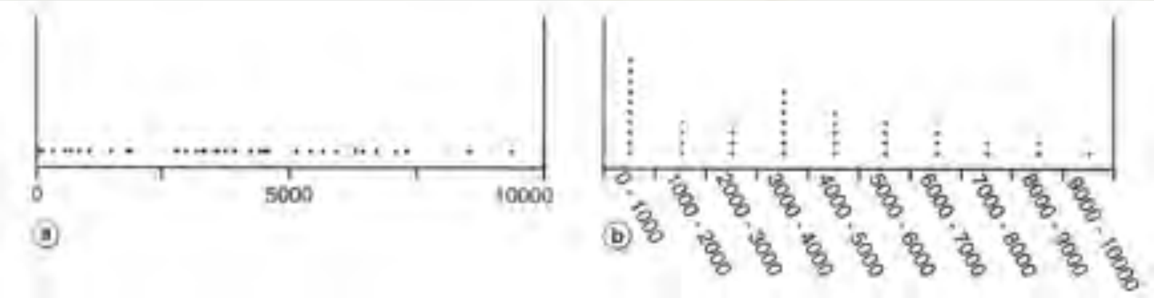
Data classification

Graphical methods of classification

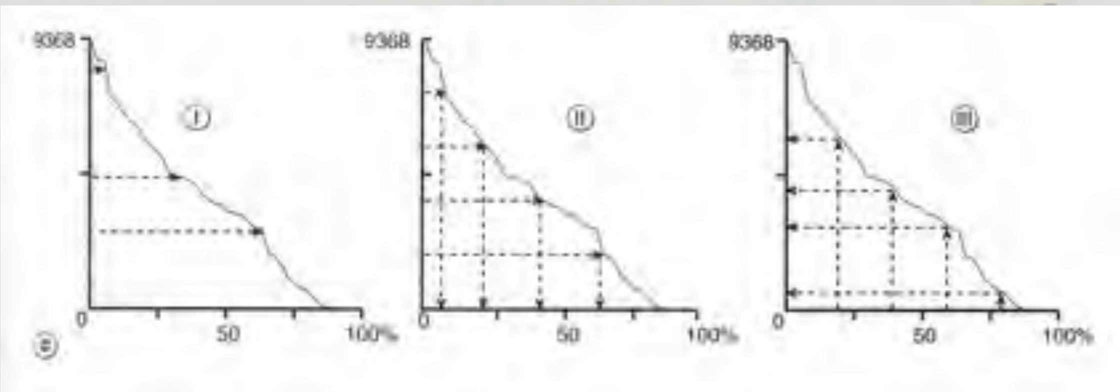
Break points – separate classes at points of discontinuities in the observation series



Frequency diagram – plot all frequencies, find discontinuities

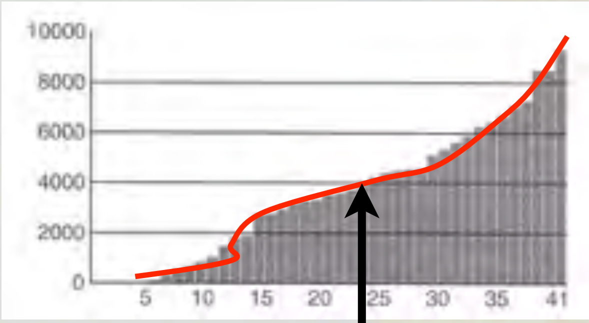


Cumulative frequency diagram – plot the added frequencies

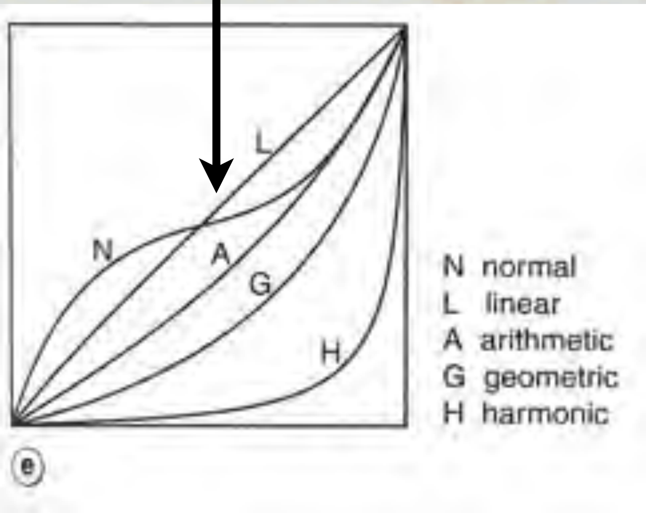


Data classification

Mathematical methods of classification

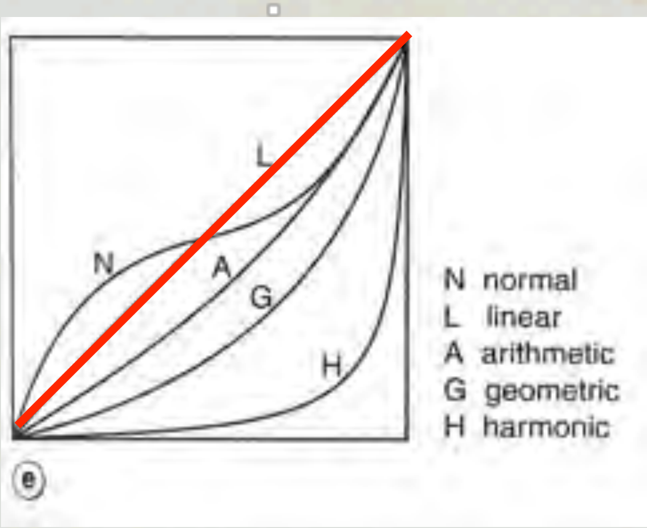


All methods draw a curve on top of observation series:
the classification method is chosen according to this curve.



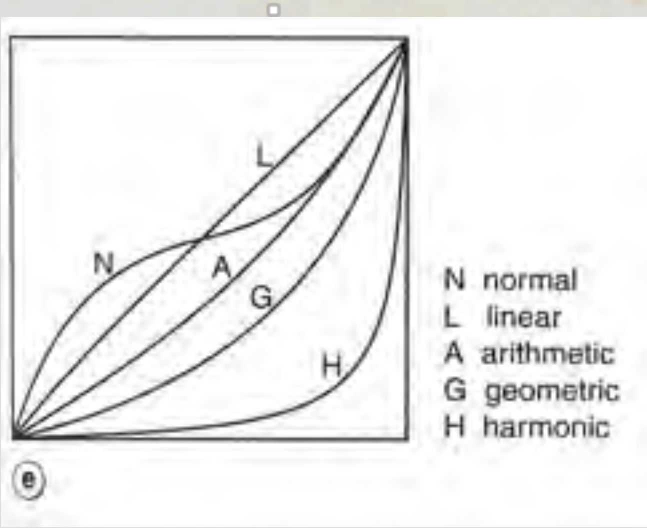
Data classification

Linear curve L – Equal steps classification – equal width for all classes



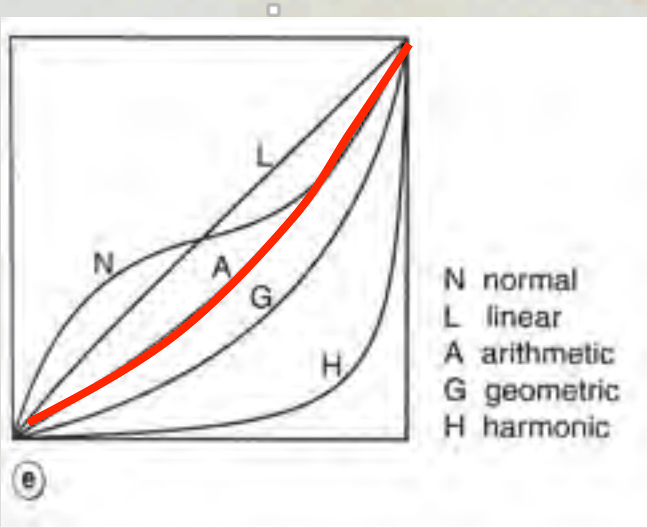
Data classification

Quantiles – splits the number of observations proportionally over all classes. 4 classes: quartiles, 5 classes: quintiles, etc.



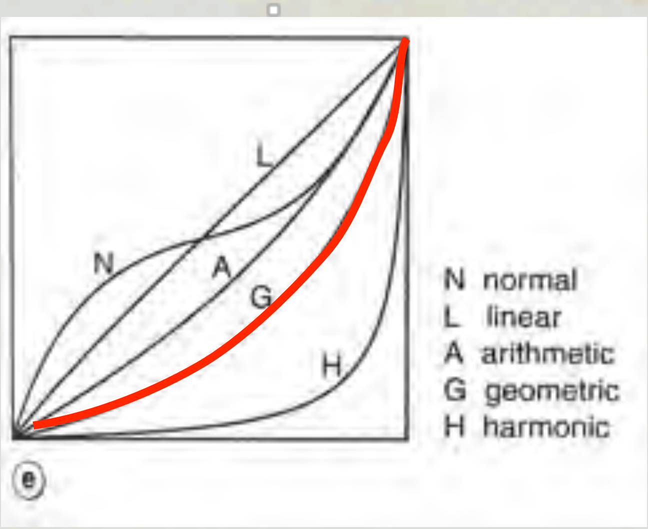
Data classification

Arithmetic series – series a_1, a_2, a_3, a_4,..., where, $a_{n+1}=a_n+c$, c=const, curve A



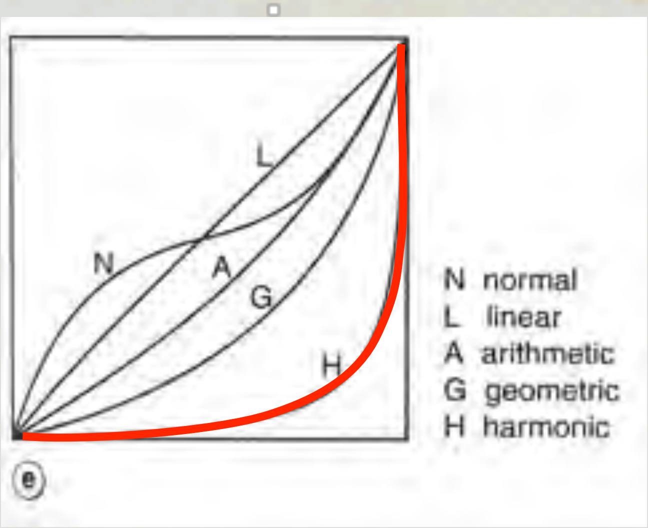
Data classification

Geometric series – series a_1, a_2, a_3, a_4,..., where $a_{n+1}=a_n*c$, c=const, curve G



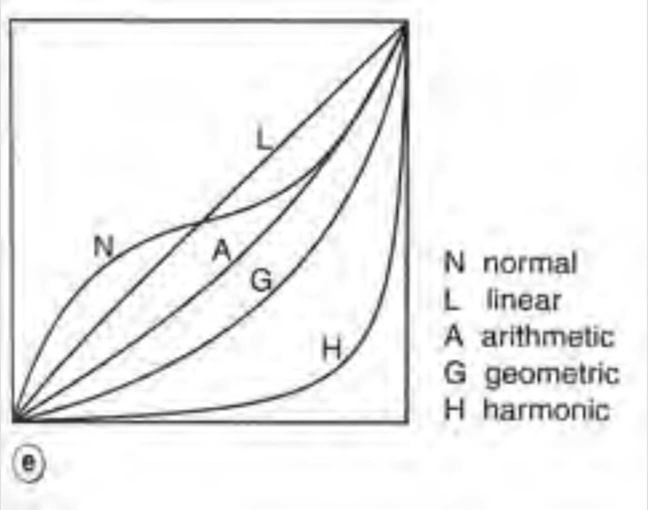
Data classification

Harmonic series – reciprocal values of the attribute form an arithmetic series - series a_1, a_2, a_3, a_4,..., where, $1/a_{n+1}=1/a_n + c$, c=const,
curve H - gives a good classification of small values



Data classification

Nested means:
1. calculate average of all values, a, set a as one class boundary,
2. calculate average of all values<a and all values>a, set these two averages as class boundaries,
3. etc. until you reach the desired number of classes (always a multiple of 2.)



Data classification

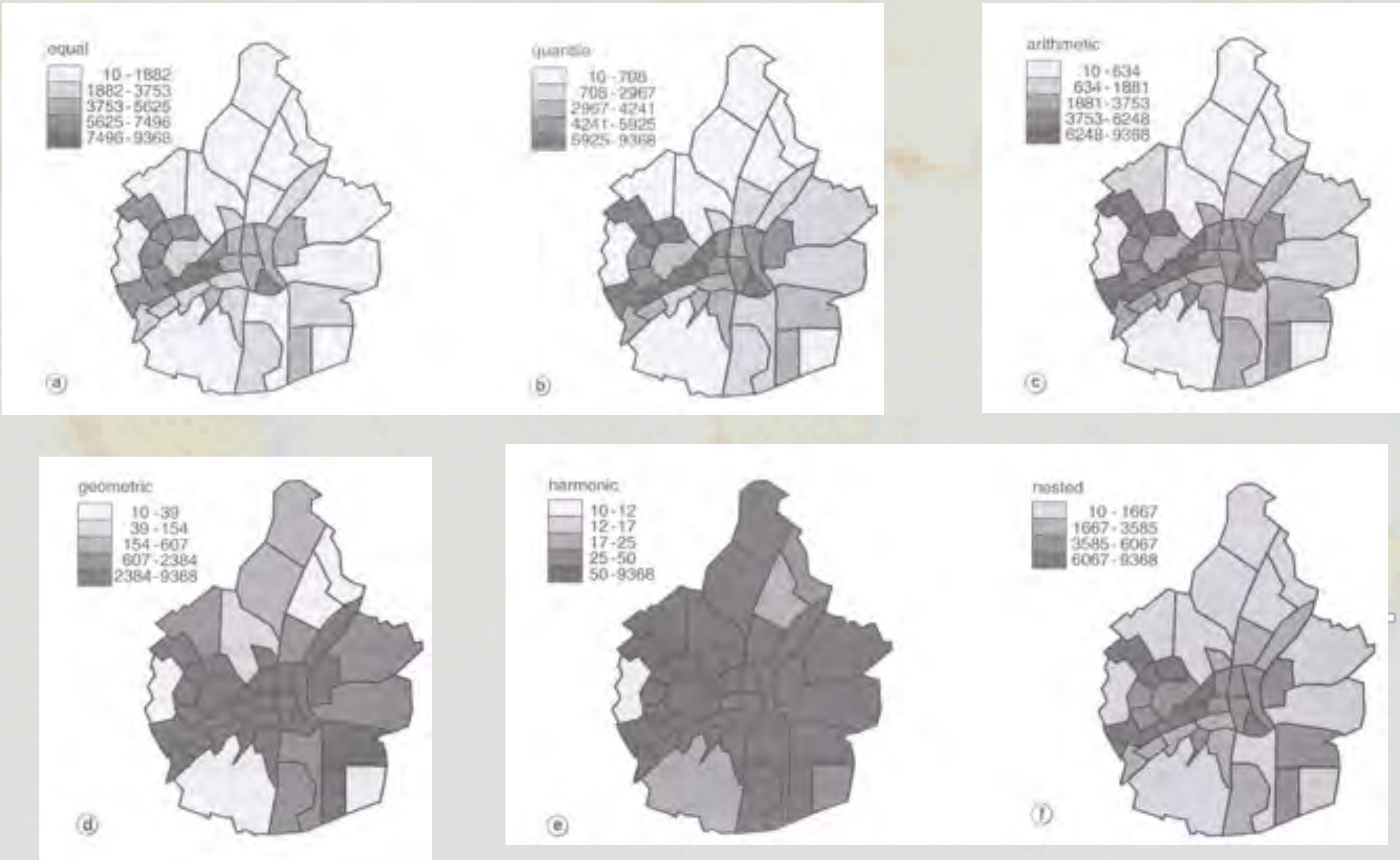
Every method results in different classifications:

equal	quantile	arithmetic	geometric	harmonic	nested
10 - 1882	10 - 708	10 - 634	10 - 39	10 - 12	10 - 1667
1882 - 3753	708 - 2967	634 - 1881	39 - 154	12 - 17	1667 - 3585
3753 - 5625	2967 - 4241	1881 - 3753	154 - 607	17 - 25	3585 - 6067
5625 - 7496	4241 - 5925	3753 - 6248	607 - 2384	25 - 50	6067 - 9368
7496 - 9368	5925 - 9368	6248 - 9368	2384 - 9368	50 - 9368	

(d)

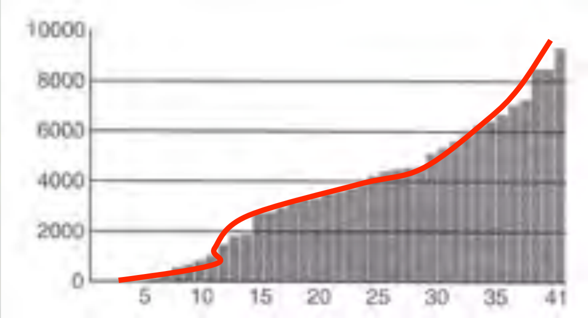
Data classification

Every method results in a different map:

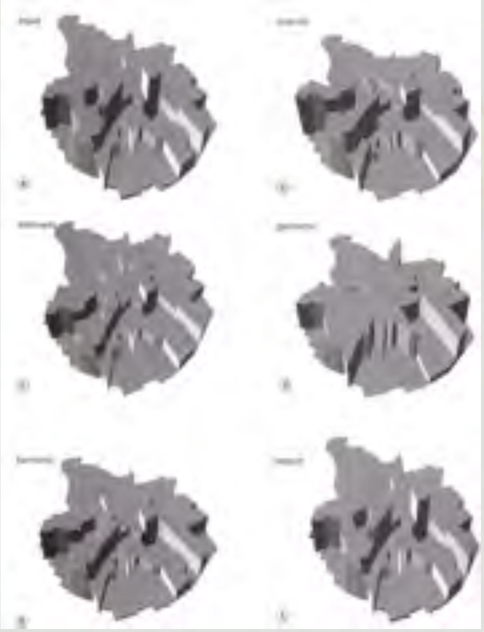


Data classification

So, which classification method is the best one?
- the one that has the curve that best fits the observation series

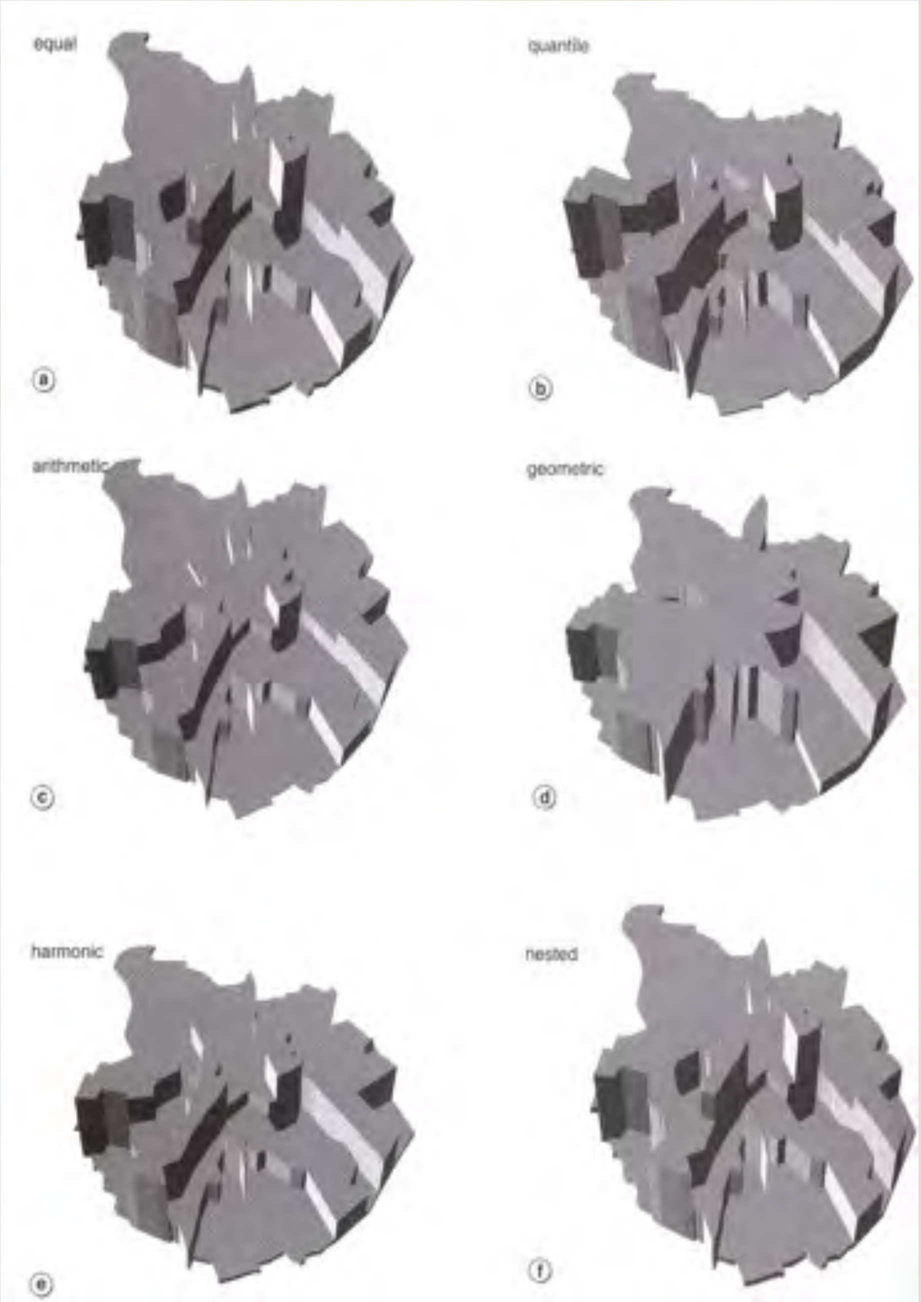


- or the one which produces the best-fitting statistical surface to the original statistical Surface:



Data classification

Classified surfaces



Original surface

Cartographical data analysis

Goal: to assess the characteristics of the components of the information and deciding which graphic variables to use for the visualisation.

Step 1: find the common denominator for all the data elements/ categories selected for representation → The title of the map

Land	Apples	Pears	Prunes	Cherries	Other	Total
Saarland	30	10	20	10	10	80
Schleswig-Holstein	120	20	20	20	20	200
Hessen	130	20	40	30	10	230
Rheinland-Pfalz	120	30	40	40	10	240
Bavaria	210	40	50	30	10	340
Nordrhein-Westfalen	280	60	40	40	20	440
Lower Saxony	390	40	40	30	10	510
Baden-Württemberg	900	160	100	30	20	1210

Fruit production in Germany in 1967

Cartographical data analysis

Step 2: assess the data variables that vary from one data element to another.
Example: soil map – the geographical location of each sample site, the various soil units

Step 3: assess the measurement scale of these variables, the range of the data and the length of variables (= the number of classes/ categories).

Step 4: assess the information hierarchy – which aspects are the most important ones, which are the least important, what data categories come in-between and in which order? Translate the information hierarchy into graphical hierarchy.

Step 5: construct a preliminary visualisation – shows trends, patterns, etc. that one should stick to during the actual mapping. Transformations are applied to this visualisation, depending on the audience and communication objectives.

Mapping methods

Mapping methods = standardised ways of applying the graphic variables for rendering information components. They take into account:
- measurement scale
- nature of distribution of objects
- continuous/discrete distribution
- smooth boundaries or not

The nine most common mapping methods:
- chorocromatic maps or mosaic maps,
- choropleth maps,
- isoline maps,
- nominal point maps,
- absolute proportional maps,
- diagram maps,
- dot maps,
- flow line maps and
- statistical surfaces.

Mapping methods

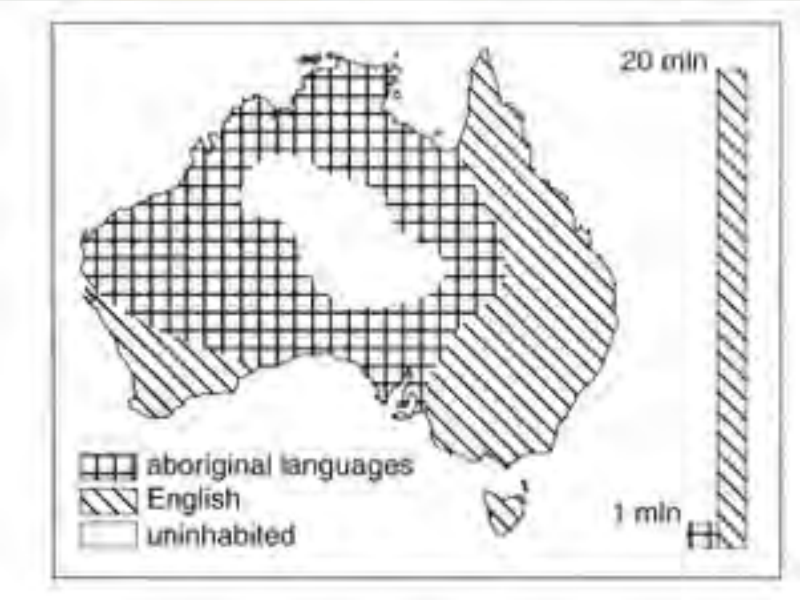
Chorocromatic or mosaic maps

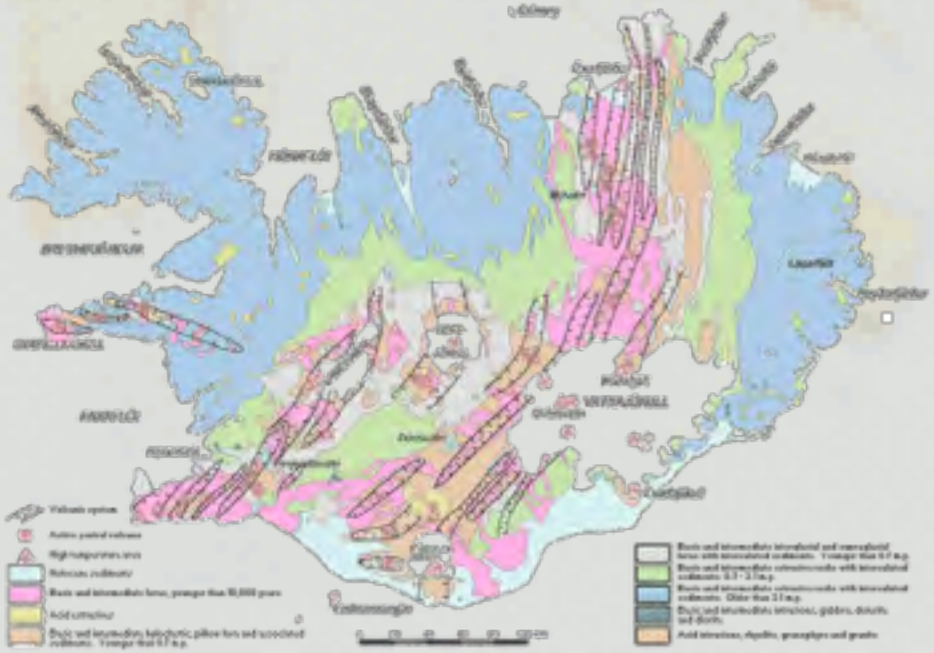
Greek: choros = area, chroma = colour

↓

Chorocromatic maps:

- rendering nominal data (qualitative) with the use of colours,
- or black and white patterns.





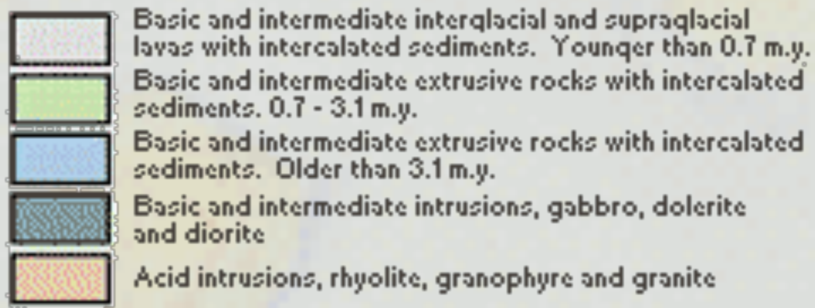
Mapping methods

Important: only nominal (qualitative) differences shown!

↓

NO hierarchy, NO order!

Different colours



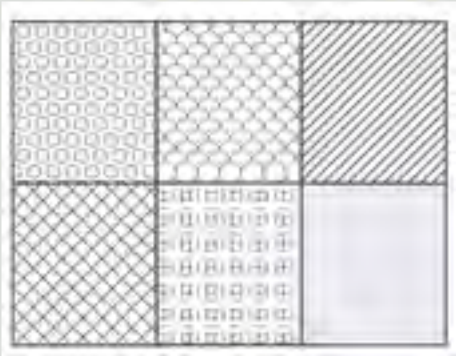
↓

But: problems with perception:
- psychological values
- perception of one colour affected by surrounding colours
- saturated colours only for small areas (no domination)

Patterns

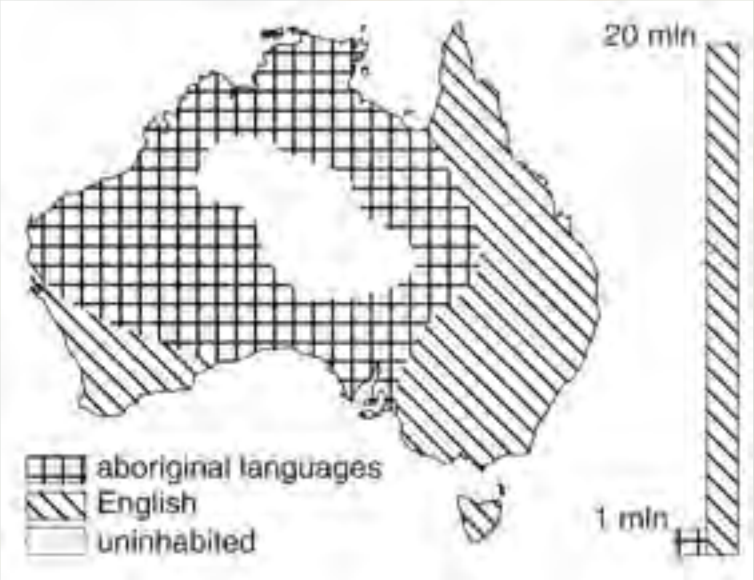
↓

But:
- they have to be easily discernible one from another
- they have to be comparable (no domination) -> same grey value



Mapping methods

Influence of the area sizes, when non-area related phenomena are mapped:



English speaking communities cover the same area as the aboriginal speaking communities.

↓

Number of English speakers = number of aboriginal speakers

→

Not true! Outback is much more sparsely populated than the coastal area with predominately English speakers.

↑

Correction: add a histogram with the number of speakers.

Mapping methods

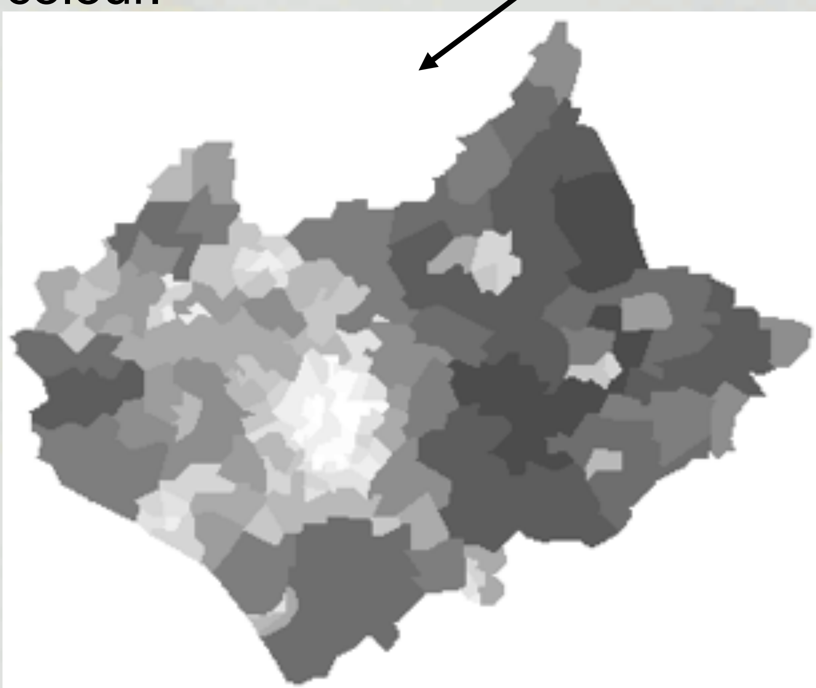
Choropleth maps

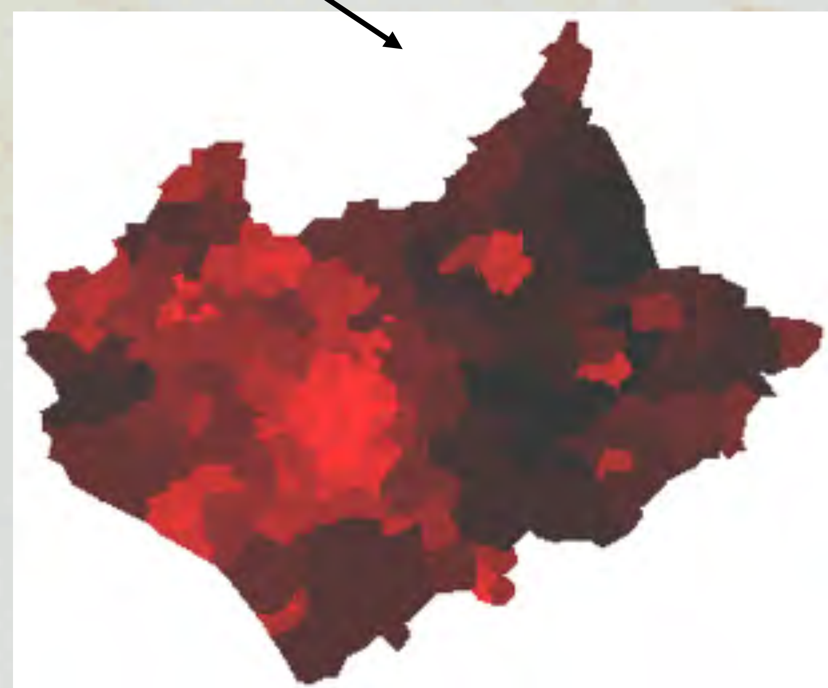
Greek: choros = area, plethos = value

↓

Choropleth maps:

- rendering discrete values (ordinal, interval, ratio data)
- values calculated for areas and displayed as a stepped statistical surface
- using differences in lightness (grey value) or saturation (chroma) of a colour.





Mapping methods

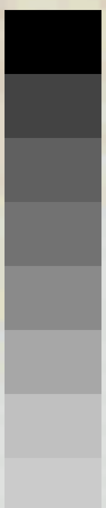
Differences in grey value / saturation

↓

Differences in the intensity of the phenomenon: if correctly applied – percentage/density that is twice as high as another percentage/ density, is represented by a twice as dark grey value.

→ Hierarchy + order

↓



Dark values: high intensity/density of the phenomenon

Light values: low intensity/density of the phenomenon

Mapping methods

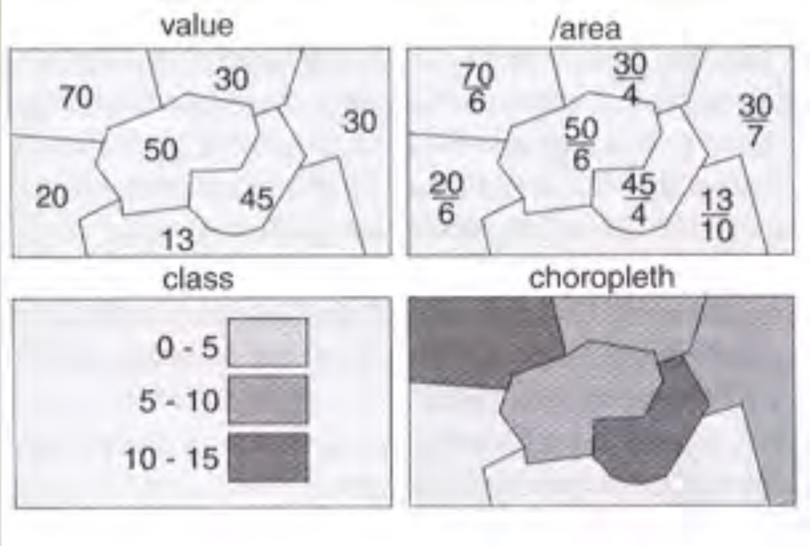
Two main types of choropleths

Density maps:
- ratio in which the areas covered are in the denominator

Non-area related ratio maps

Important difference for map-use: visual impression of a choropleth is affected by both the tint and the size of areas.

Production procedure for both types:
1. absolute value
2. put the absolute values into perspective
3. classification (max. 7 classes)
4. assigning grey values to classes



Mapping methods

Classification: simplification of the image.

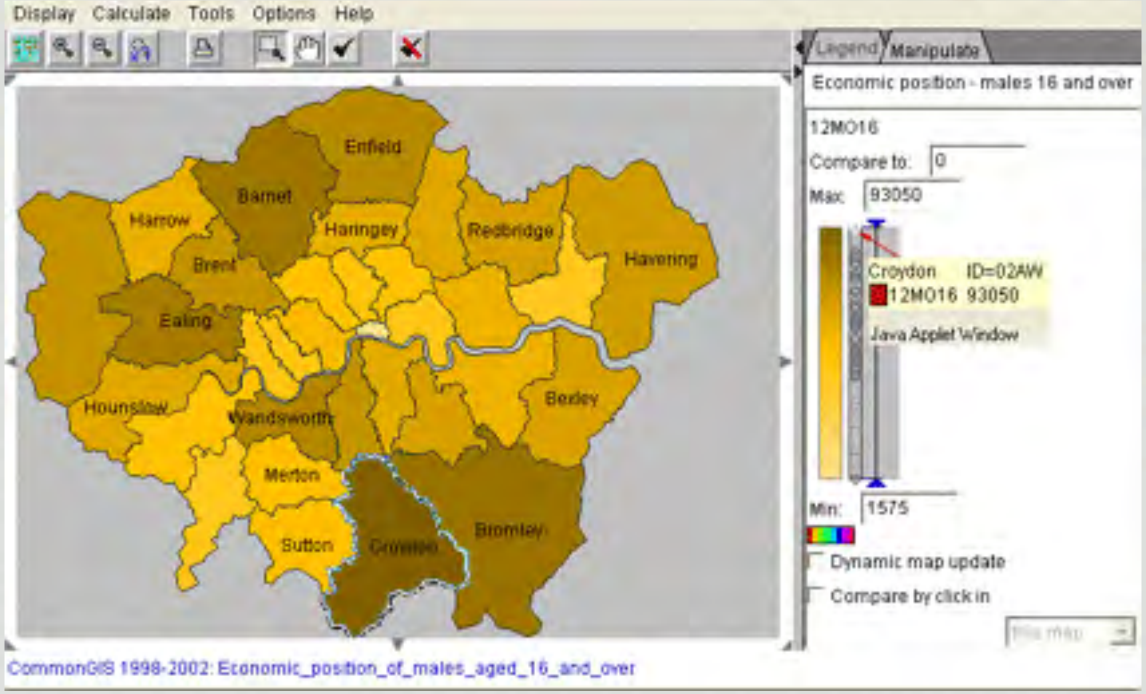
Condition:

minimal difference within classes + maximum difference between classes.

↓

If the condition can't be met:

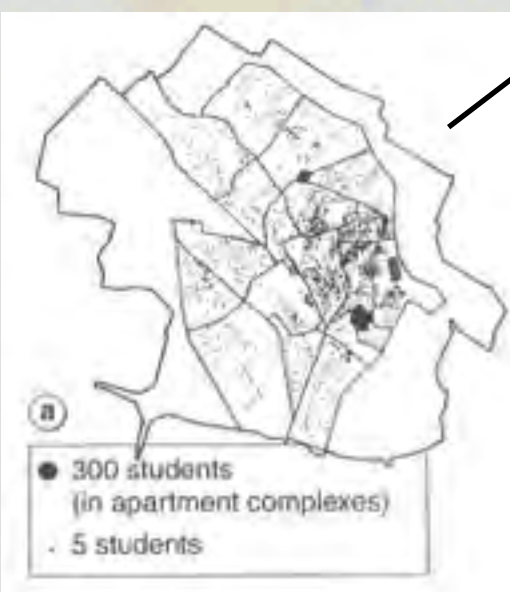
unclassified choropleth map with a continuous scale



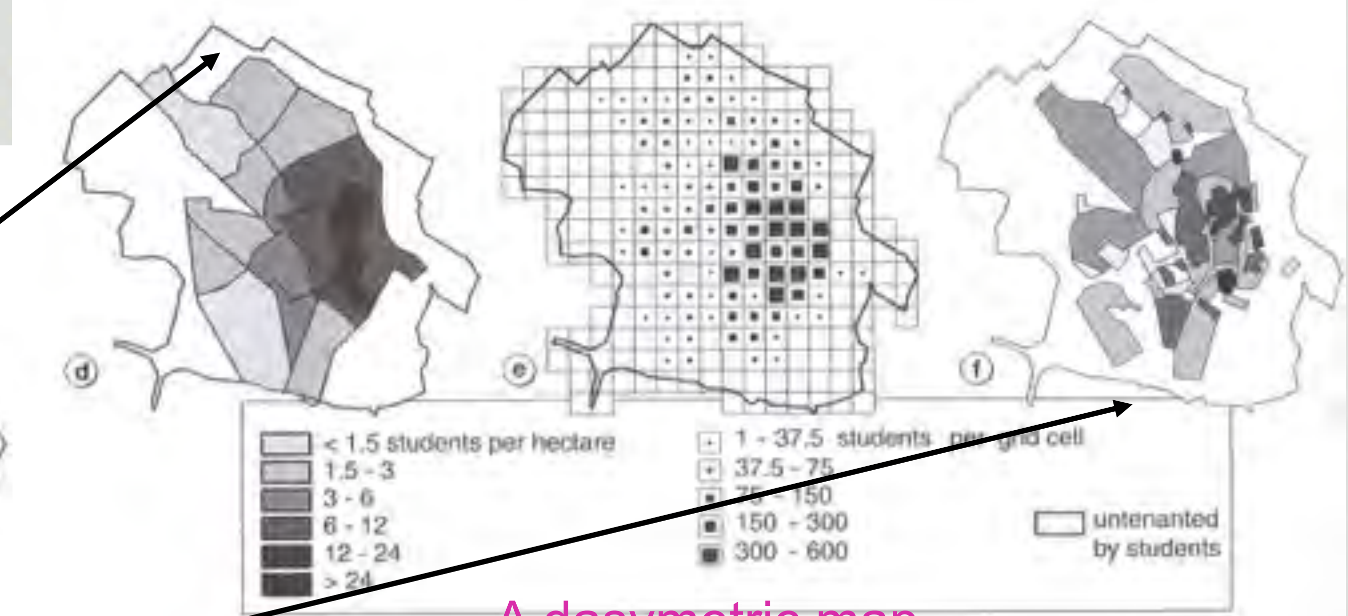
Mapping methods

A dasymetric map:
A choropleth map where area boundaries are adjusted to the occurrence of the phenomenon.
Usually: area boundaries are artificially created boundaries (I.e. administrative units or similar) that have nothing to do with the phenomenon.

A dot map of the density of student population in Utrecht

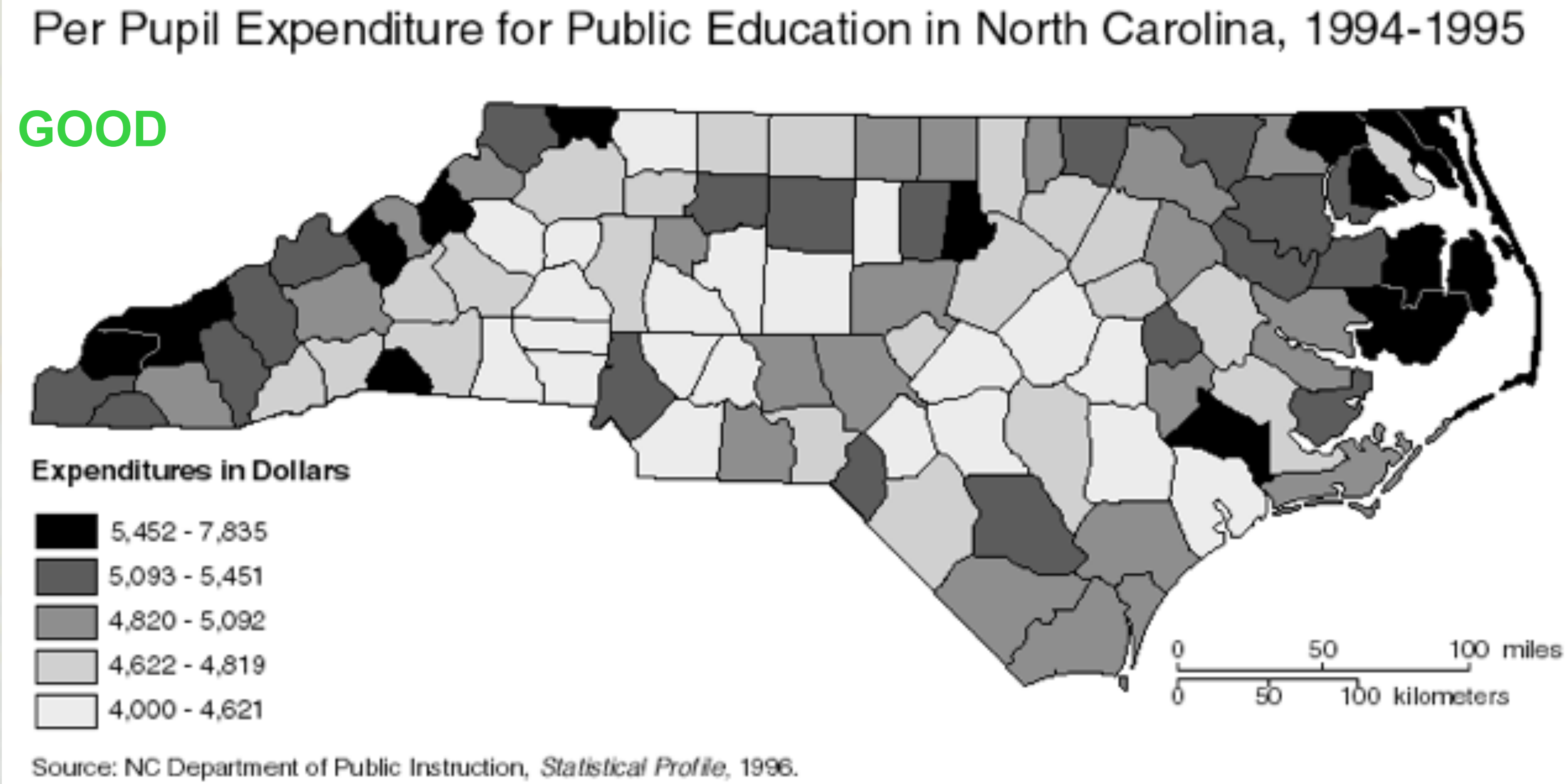


A choropleth with statistical demarcation areas



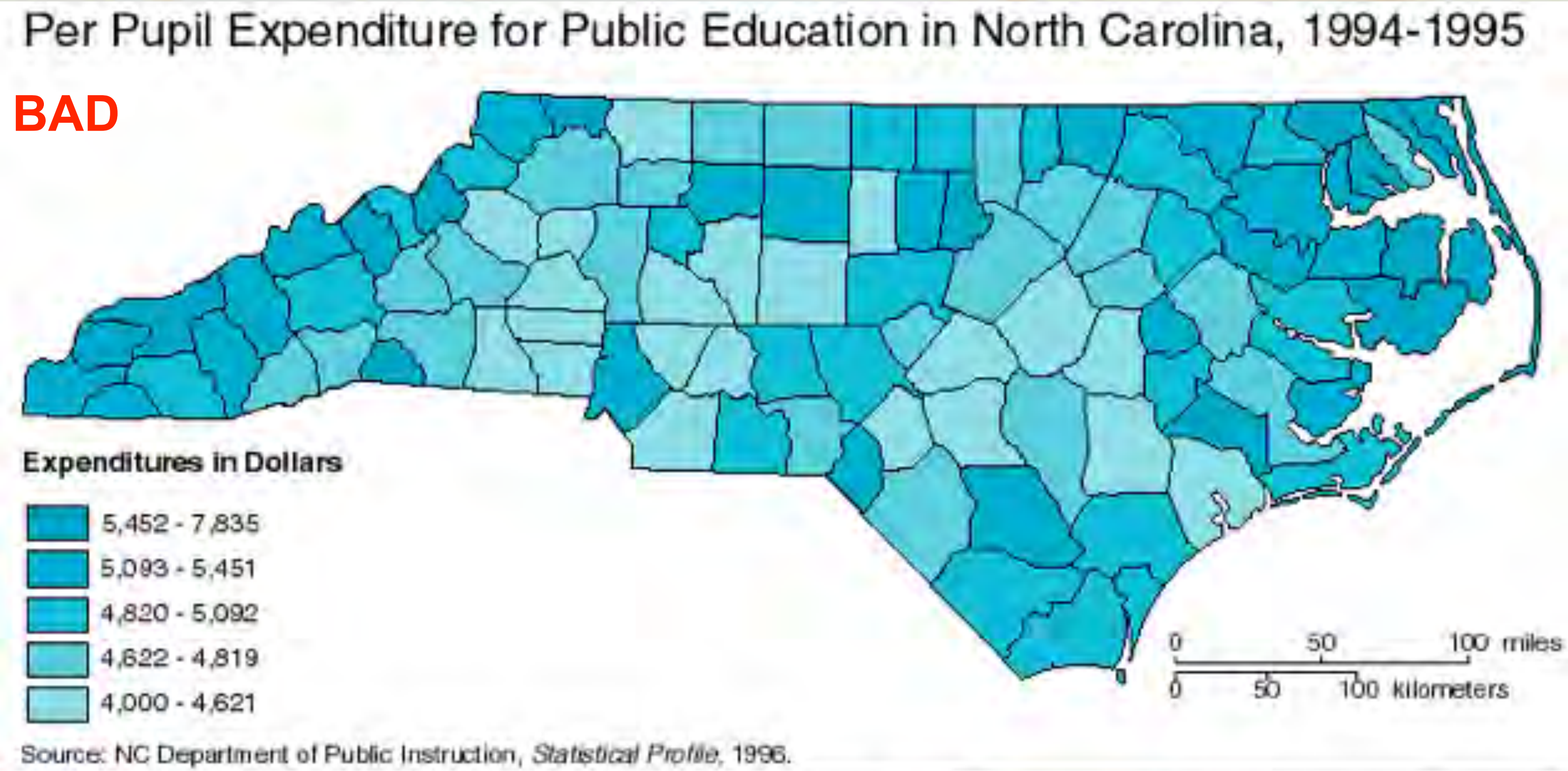
A dasymetric map

Mapping methods



5 Classes can be represented sufficiently in black and white as well as by a single hue. The top class must be black and there must be at least 15% difference at the lower end of the value/chroma scale, at least 20% difference in the mid range and at least 25% for dark ranges. Percentages of black used here are: 8, 24, 50. 70 and 100%. Note that when adjacent polygons have black, the dividing lines are not visible when also in black. The polygon outlines must, in this case, be shown in a light gray or white. They are shown here in light gray where 2 or more adjacent counties are filled with black.

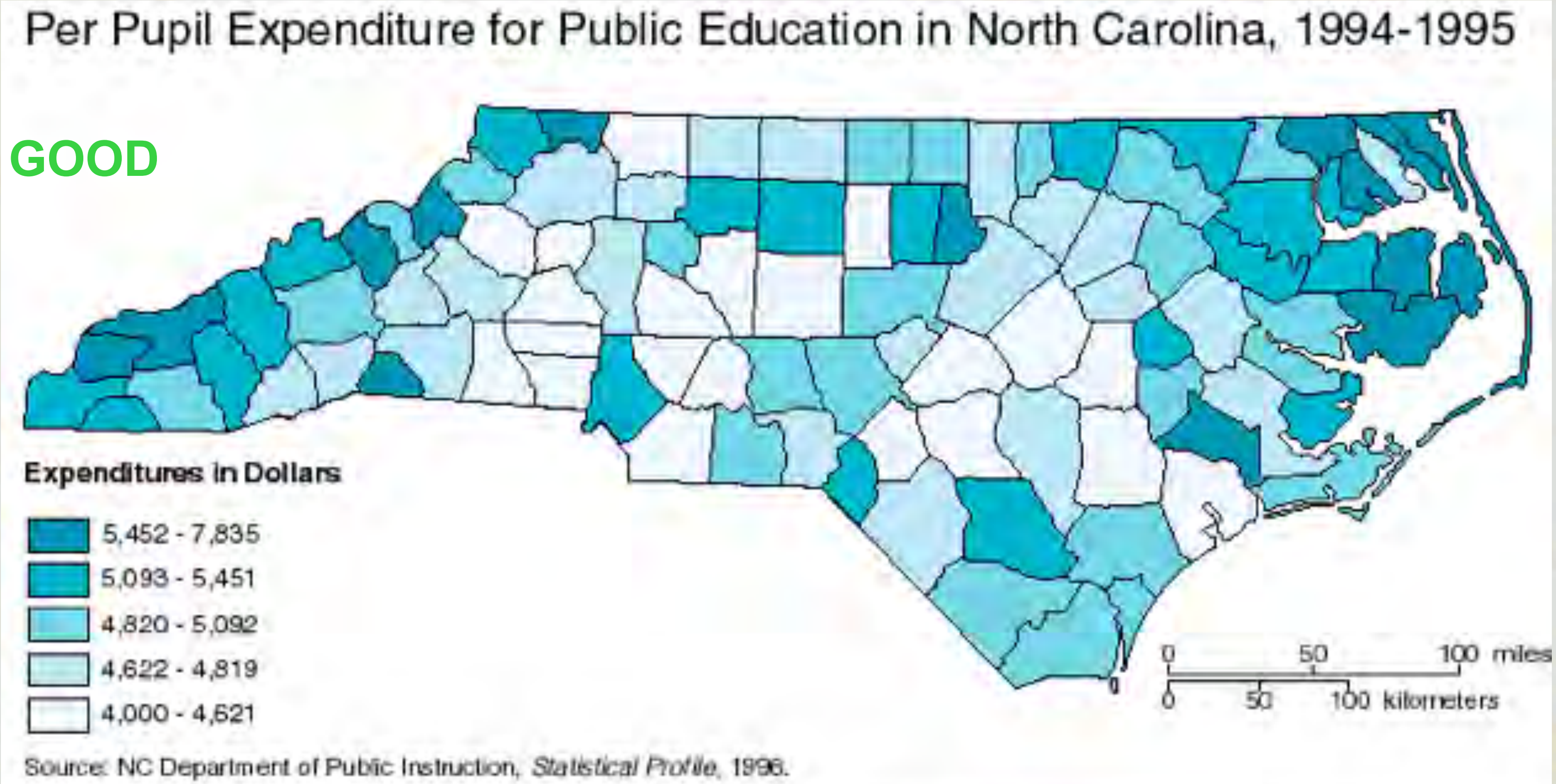
Mapping methods



In this single-hue, 5-class choropleth map, the hue cyan is used to represent education expenditures. The percentage of cyan is varied ineffectively for providing a clear distinction between each class and for highlighting the highest data quantities. The percentages used are: 100, 85, 70, 55, and 40. A difference of 15% per step is not visually distinct enough to make for an effective single-hue series. There should be 20% or more difference in chroma (the amount of hue used, expressed in percent) between the darker hues in the single-hue series. At the lighter end of the scale, the lightest hue should be quite light (e.g., 10%) with at least 15 to 20% to the next hue for effective differentiation of low chroma hues. With 4 or fewer classes, it is not necessary to vary the value (i.e., no need to use black). But with 5 classes and a single hue, decreasing the value (adding small amounts of black) for the darkest hue helps with differentiation per hue and with highlighting the highest data quantities.

Mapping methods

GOOD



Shown here is an example of a single-hue graded series for a 5-class choropleth map. Single-hue series' really work better for 4 or fewer classes but a single hue can be "stretched" to 5 distinct value/chroma steps by varying the percentage of the hue, and decreasing the value (adding black) of the hue for the highest data class. Percentages shown here are: 8, 24, 48, 80, and 100% Cyan with 10% Black added to the 100% Cyan. Percentages will vary depending on the hue used. Green works well for a 5-class single-hue graded series. Magenta, red, orange and brown can also form 5-class single hue series' but yellow would not work well.

Mapping methods

Choropleth map = the prototypical thematic map
- most commonly used for portraying socio-economic data.

Advantages:
- easy to produce and read,
- patterns are easy to recognize.

Disadvantages:
- no variability within zones,
- boundaries of zones are often not related to phenomenon.

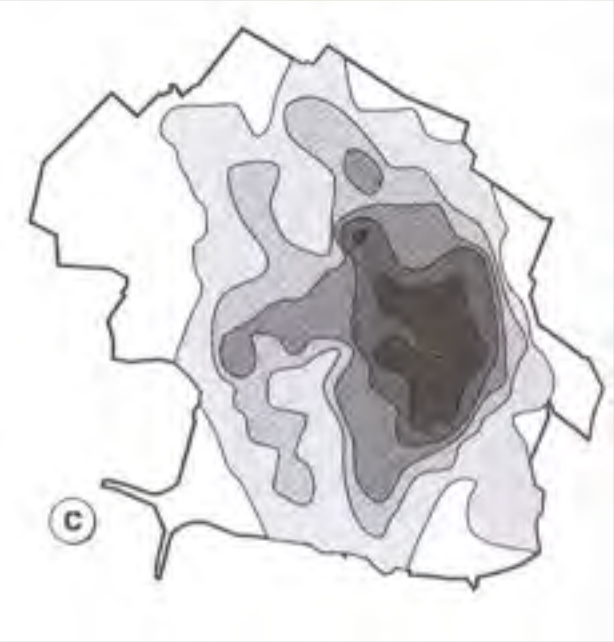
Mapping methods

Isoline maps

Greek: iso = equal

↓

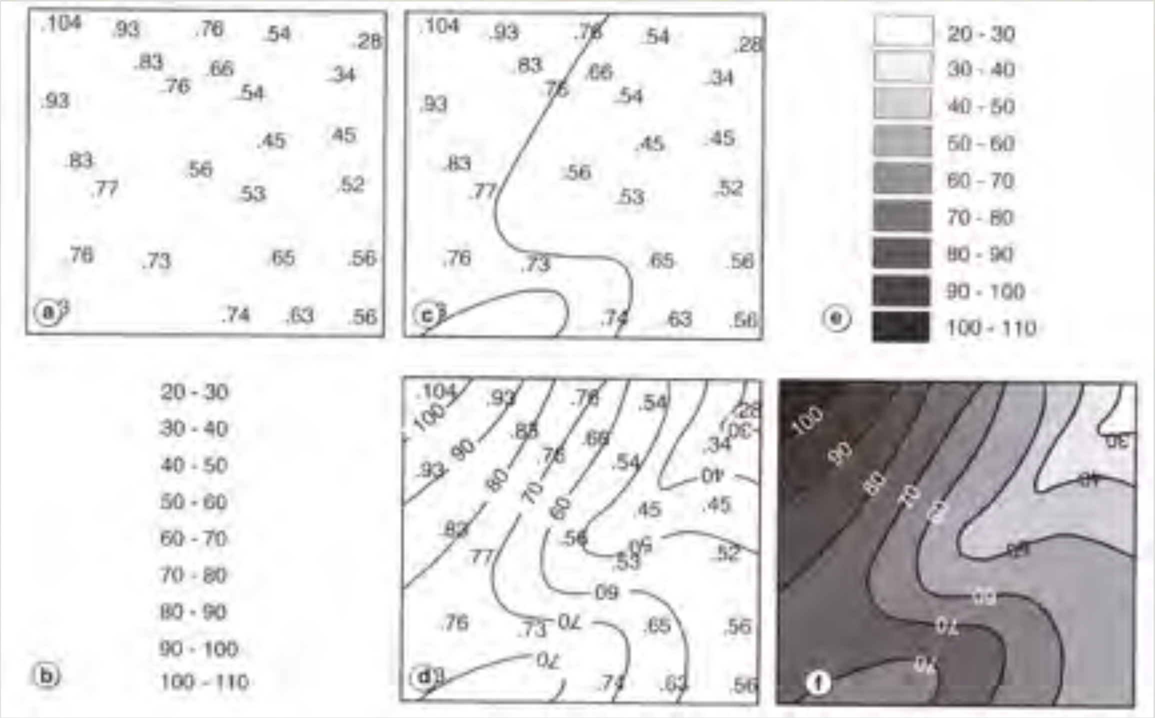
Isoline maps:
- represent continuous phenomena
- isoline = a line which connects points with an equal value
- the data: measurement values that refer to points or areas



Mapping methods

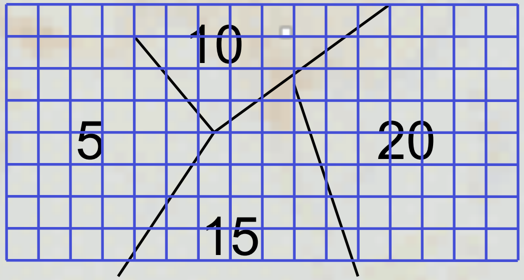
Production procedure for isoline maps with point-based data:

1. measure data in sampling points
2. categorise the data in classes
3. draw class boundaries by interpolation
   - construct the points that "have" the class boundary value
   - connect these points -> isolines
4. Add tints in-between isolines to better perceive the general trend.



Mapping methods

Production procedure for isoline maps with area-based data:
1. cover the areas with a grid and assign the appropriate values to each grid cell
2. – 4. same steps as for point-based data.



These maps are called:

Pseudo isoline maps
(Europe)

Isopleth maps
(UK)

Isoline maps show trends:
- in which direction the values of the phenomenon are increasing/ decreasing
- comparison between different phenomena and finding correlations between them.

Mapping methods

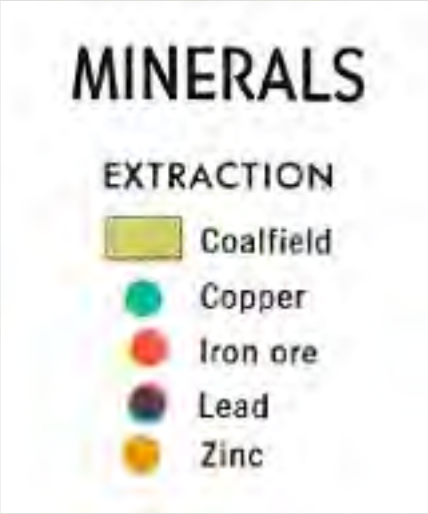
Nominal point maps

- representing nominal data valid for point locations
- by symbols, different in shape, orientation or colour.

Symbols

Figurative

Geometrical

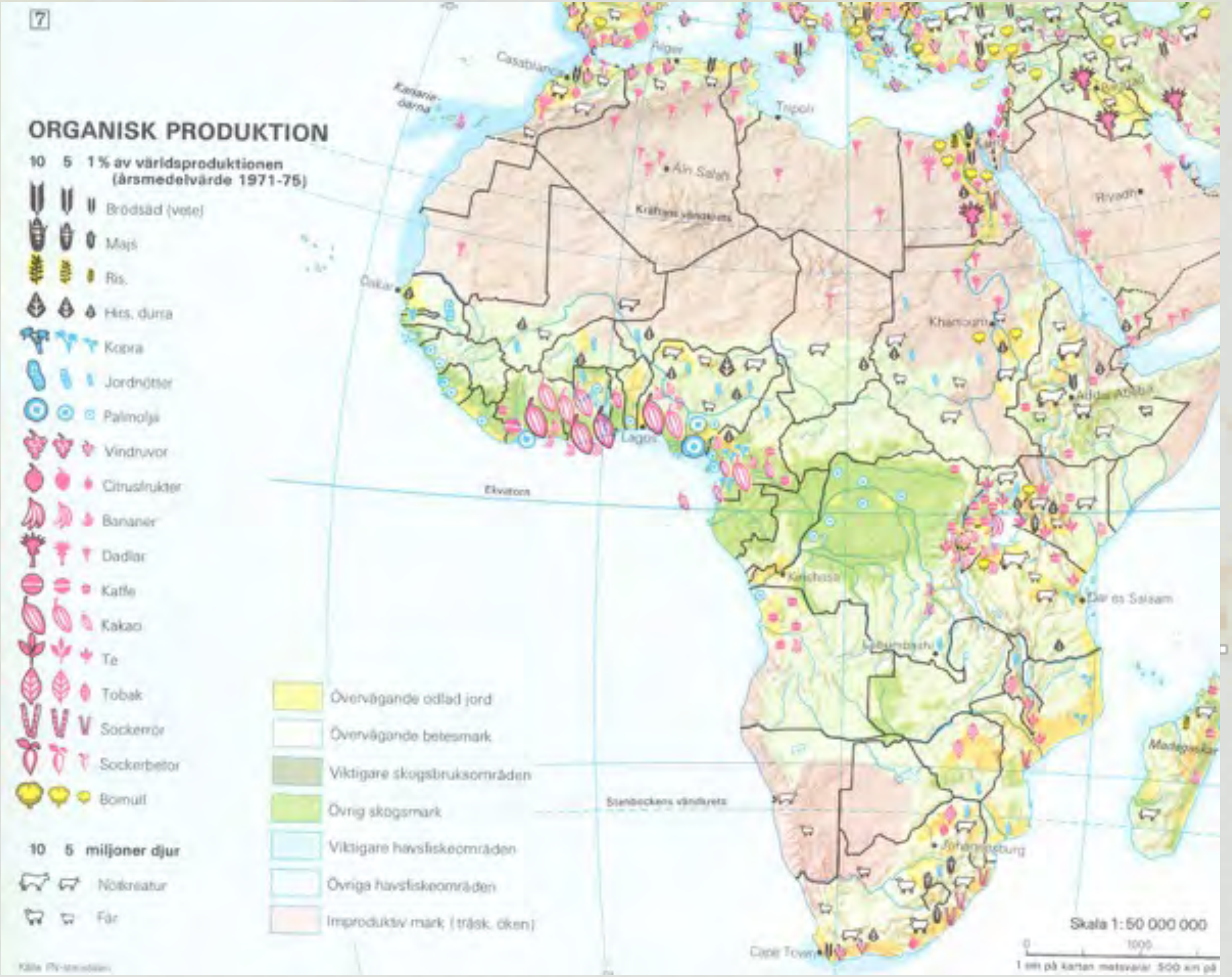


Easy recognition, difficult legibility of complex symbols

Less easy recognition, better legibility

Mapping methods

Figurative symbols



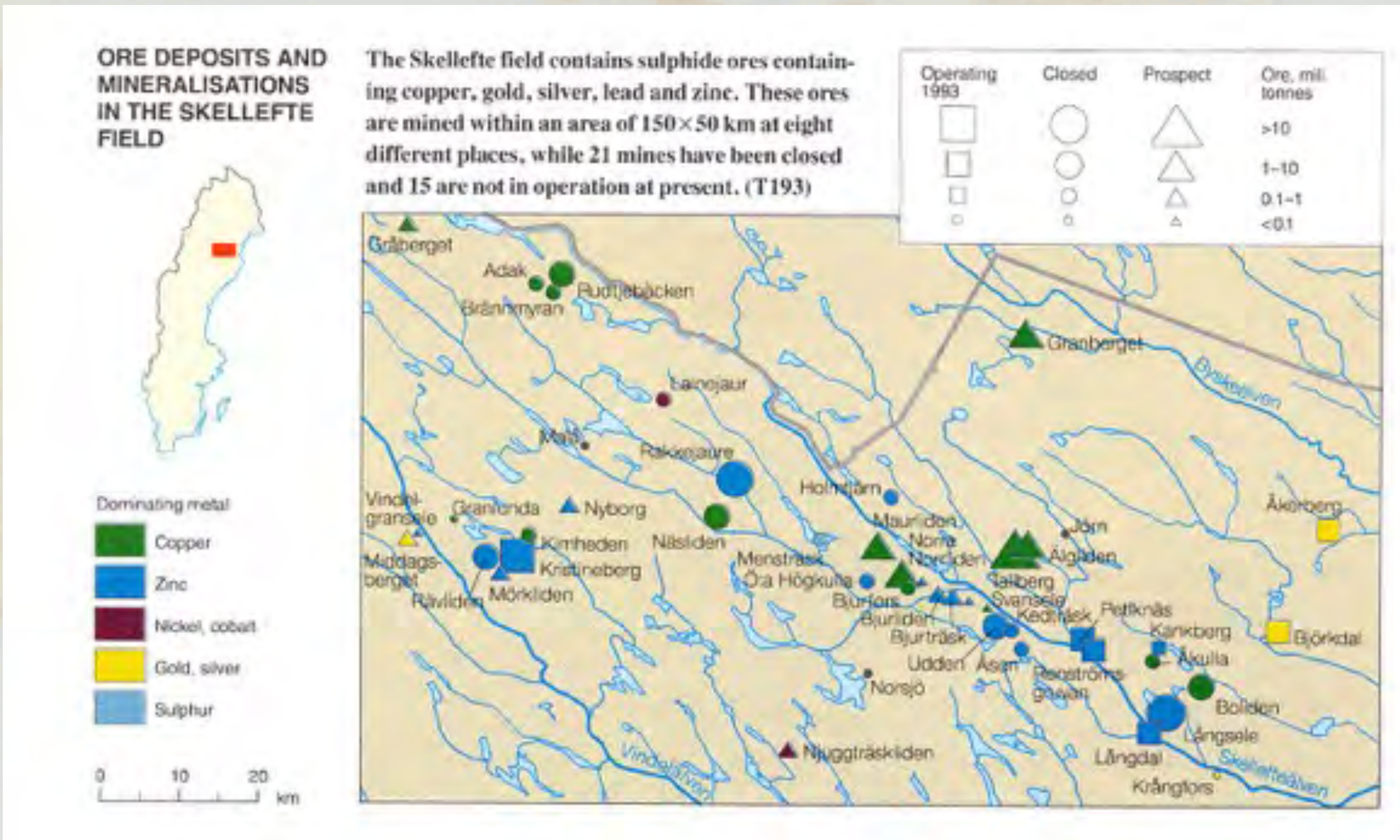
Mapping methods

Geometrical symbols

Mapping methods

Absolute proportional maps

- representing discrete absolute values valid for point locations
- by geometrical symbols, where the size of the symbol represents the value of the attribute.

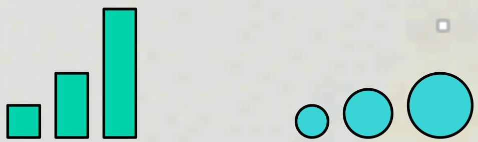


Mapping methods

Requirements:

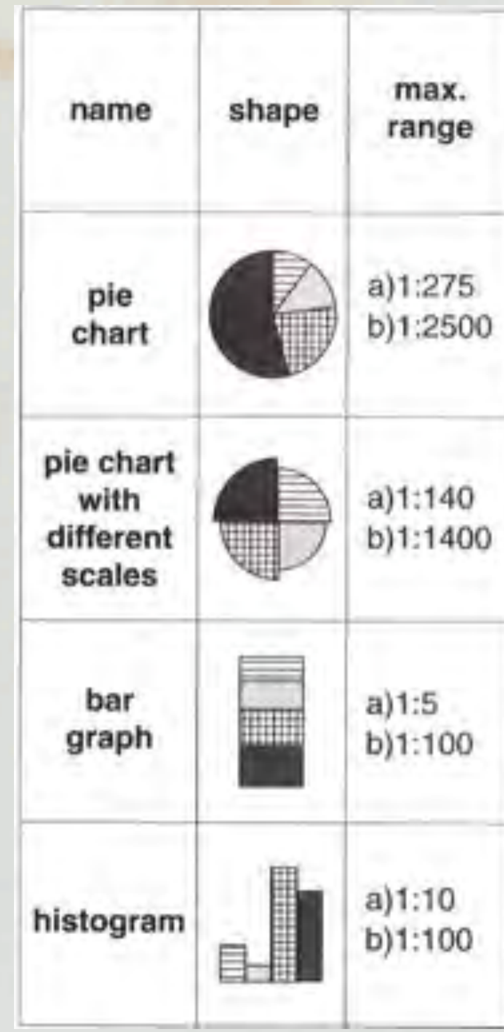
- legibility – depends on symbol density and contrast
- comparability – depends on the shape of symbols

-> easier to compare sizes of symbols that grow in one direction than circles



The range: the ratio of the highest and lowest value that can still be represented proportionally without imparing legibility.

Effectiveness of different diagrams:
(a) within each figure
(b) between figures

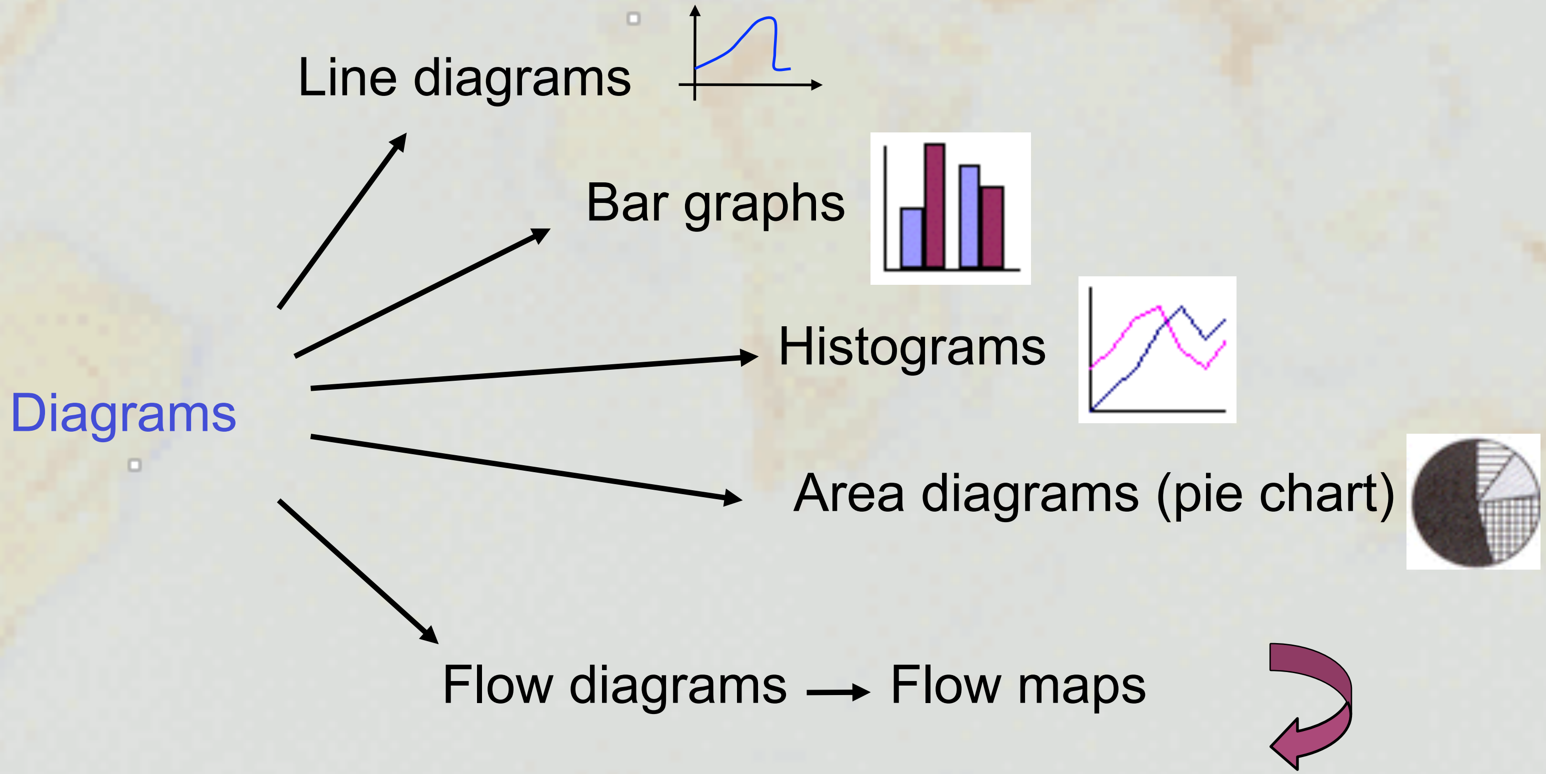


name	shape	max. range
pie chart		a)1:275 b)1:2500
pie chart with different scales		a)1:140 b)1:1400
bar graph		a)1:5 b)1:100
histogram		a)1:10 b)1:100

Mapping methods

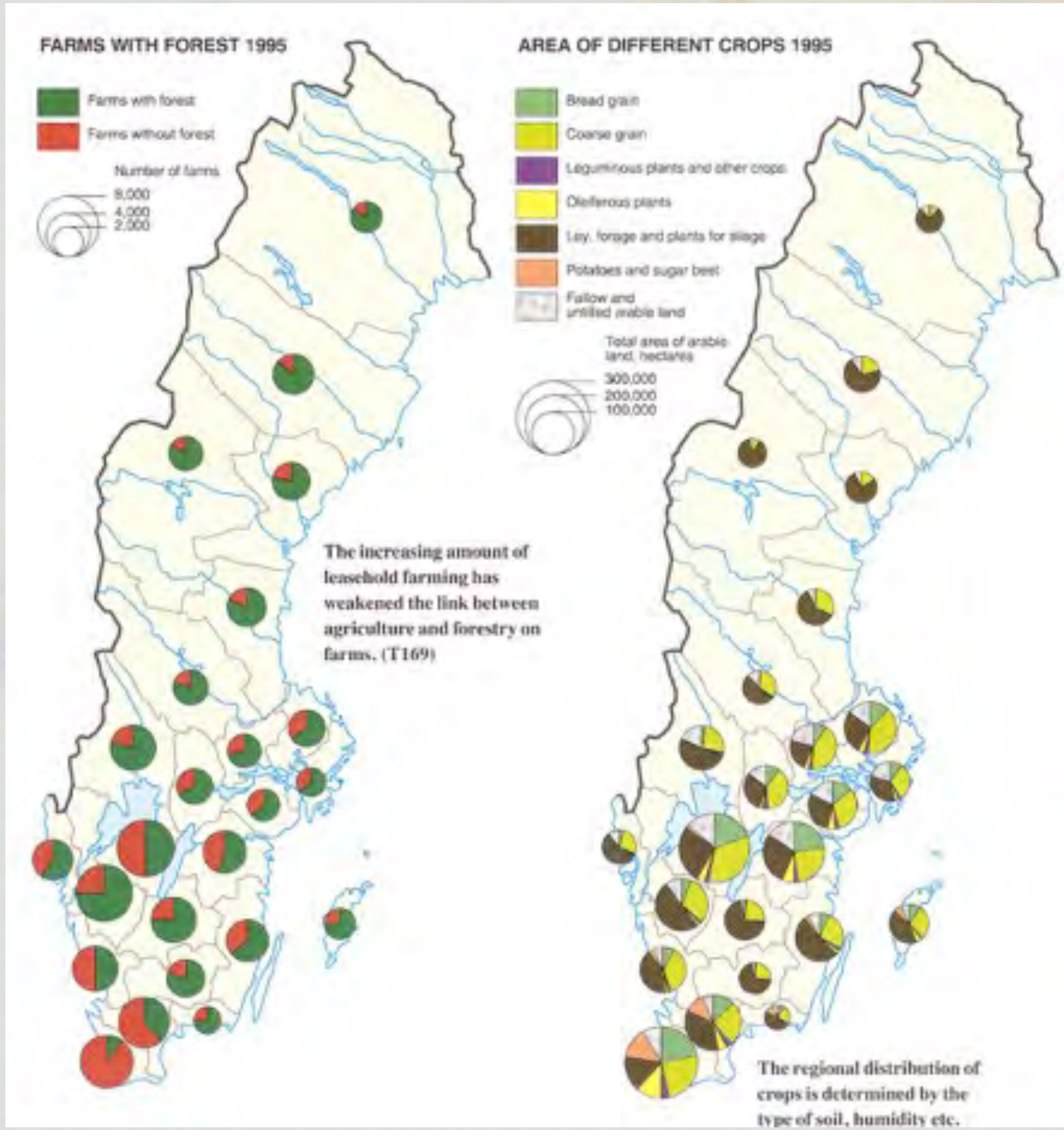
Diagram maps

- maps that contain diagrams



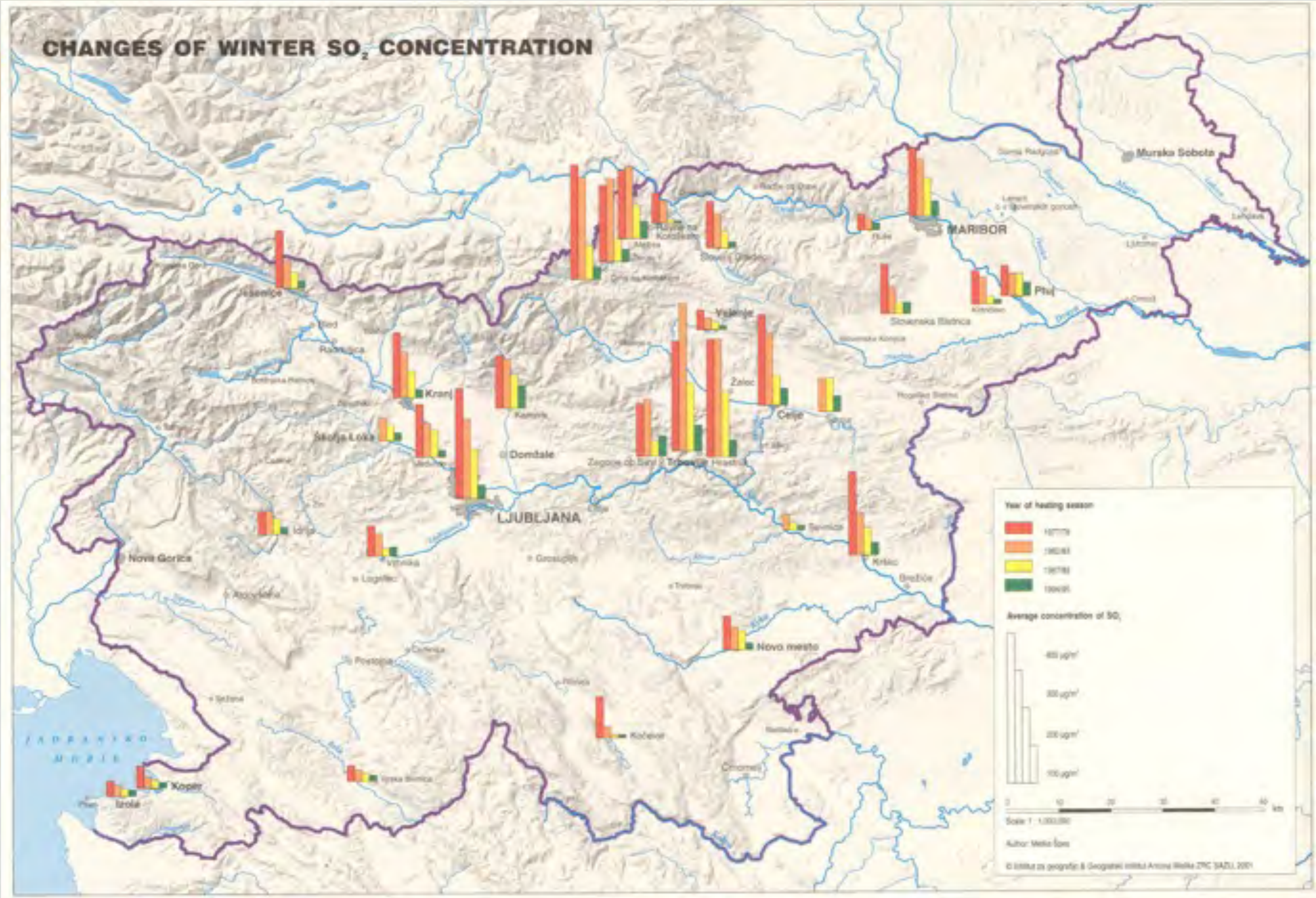
Mapping methods

Pie chart maps



Mapping methods

A bar chart map



Mapping methods

Dot maps

- a special case of proportional symbol maps
- represent point data through symbols that each shows the same quantity and is located as near the actual location as possible
- show patterns

Mapping methods

Flow maps

- simulate movement by:
  - using graphical variables that give an ordered impression
  - by showing a number of situations adjacent in time
  - by using symbols, associated with movement

arrows → the route of the movement; the volume transported; the direction of the movement

Mapping methods

Wind speed maps

Vector maps – size of the force at a specific point

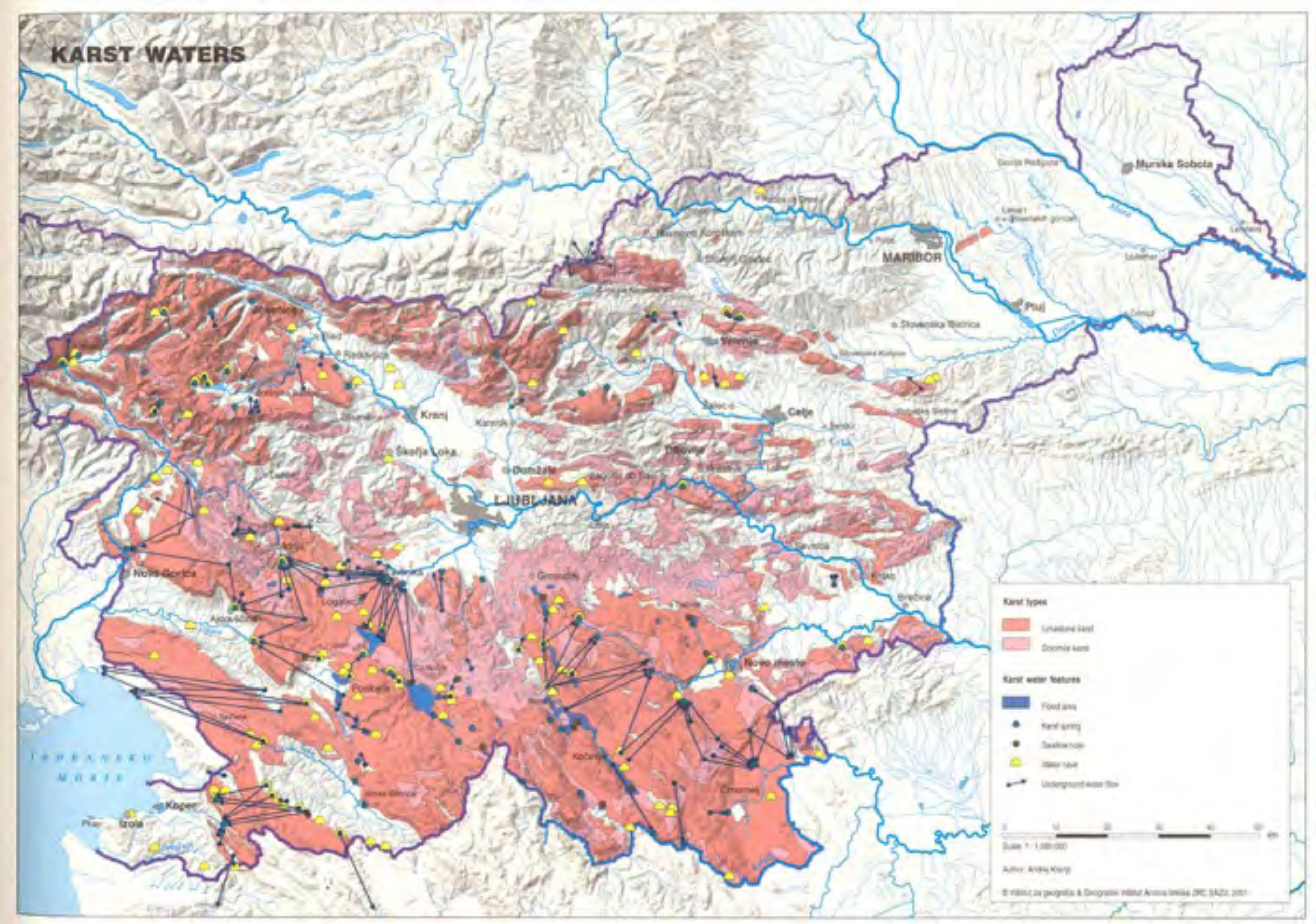
Flow maps

Flow line maps – route of the movement or transport + the size of the transport

Flow line diagrams – route of the movement + the size of the transport + the proportion of things transported by different means

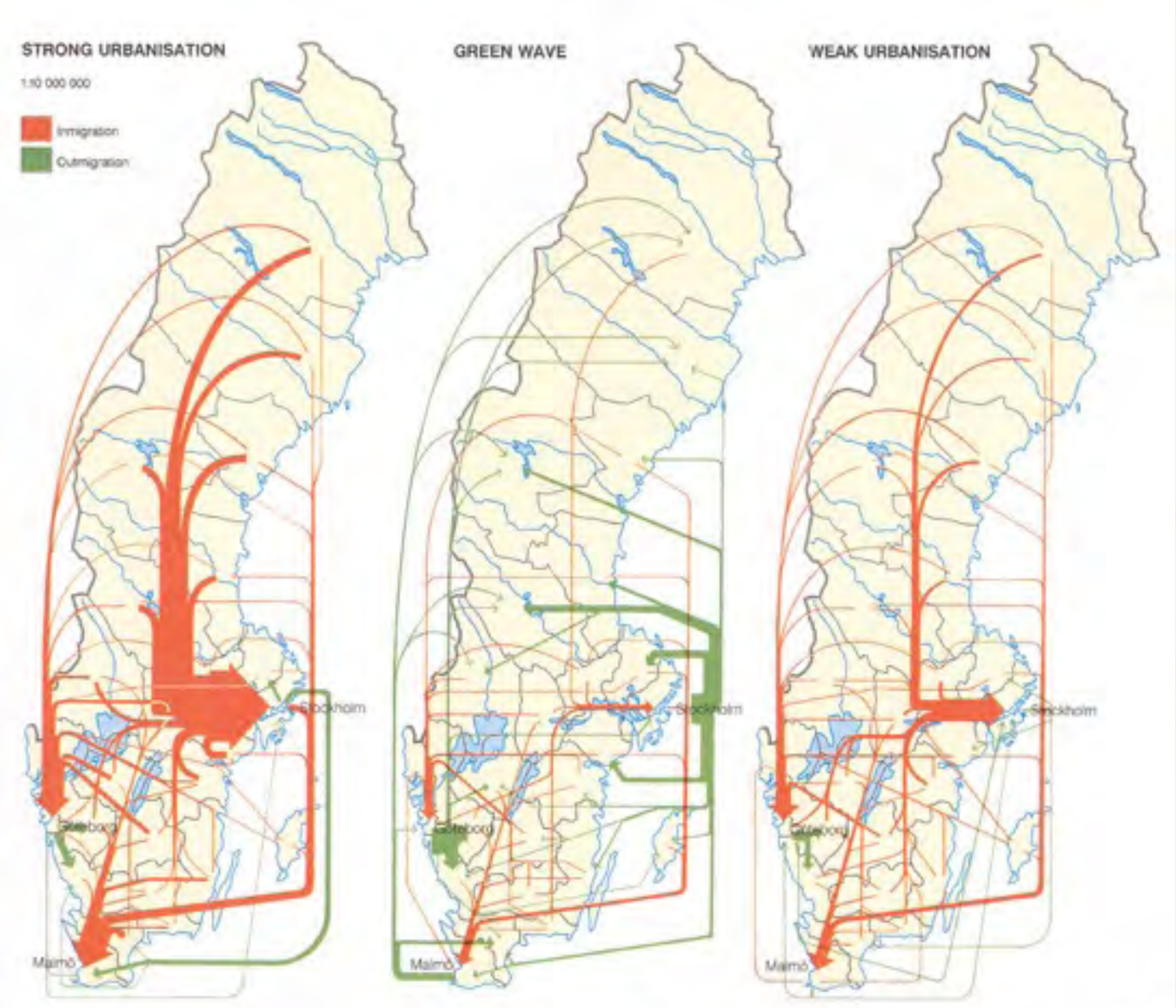
Mapping methods

A vector map



Mapping methods

Flow line maps



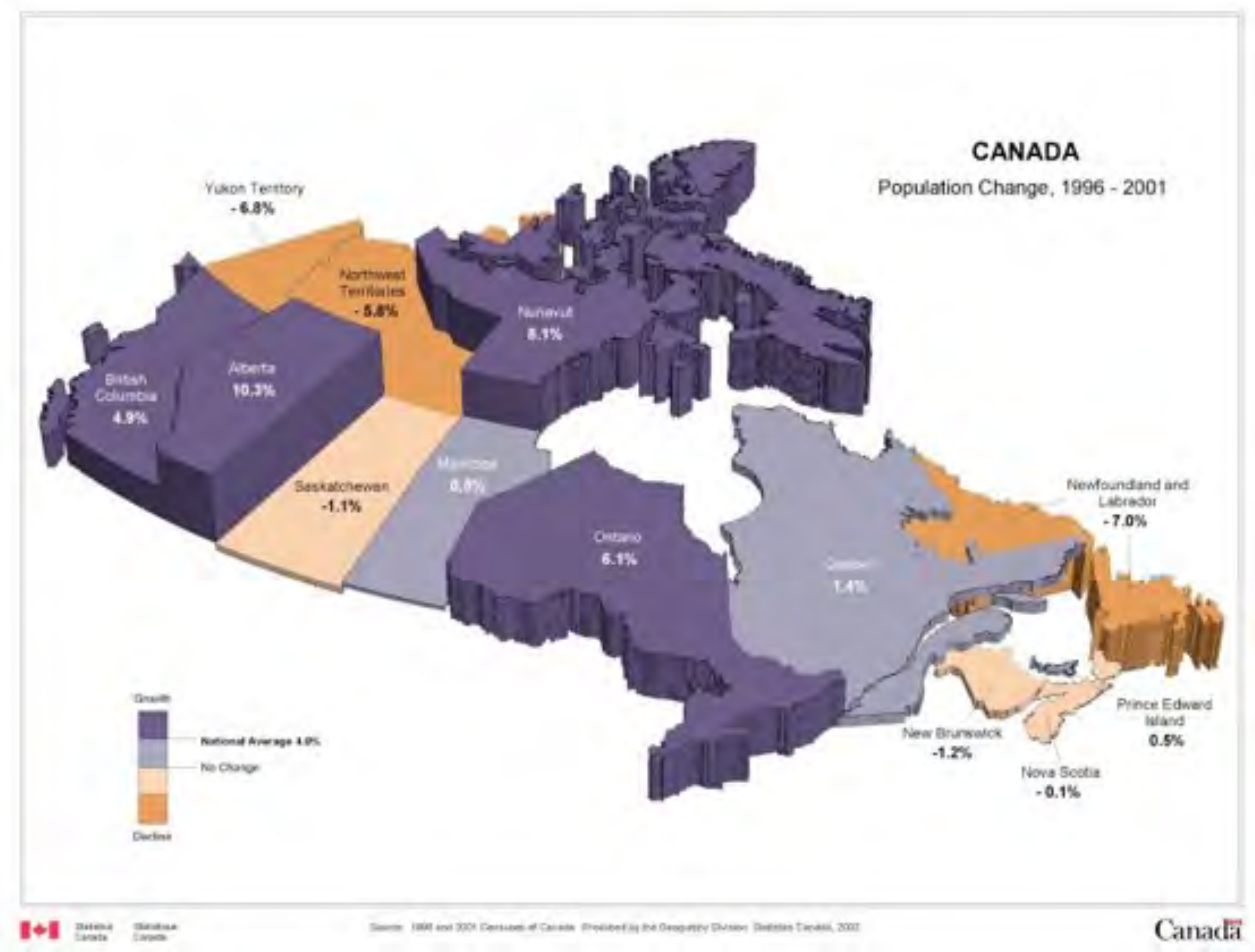
The "Green Wave" of the 1970s, which meant net outflows from the large towns. (T237)

The 1980s were characterised by slight inflows of people to the big cities. (T238)

Mapping methods

Statistical surfaces

- 3D representation of quantitative data (as in isoline/choropleth maps)

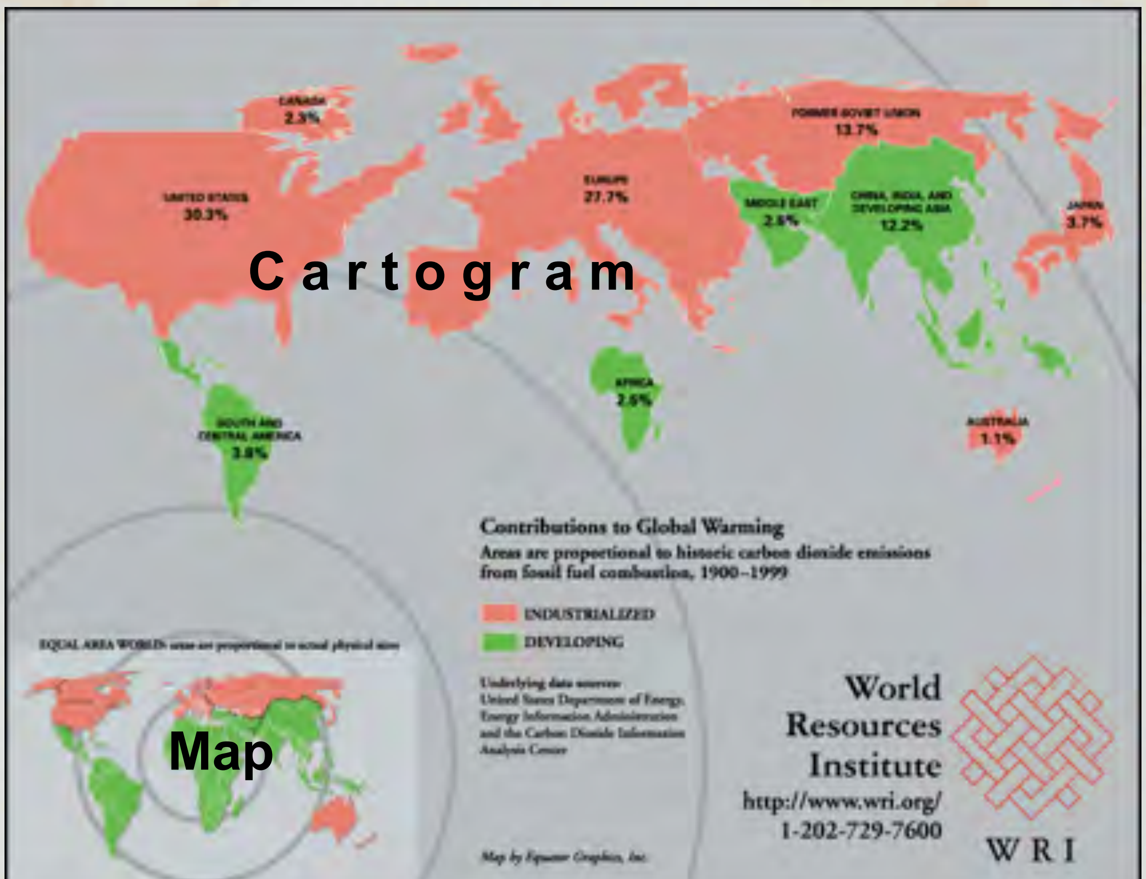


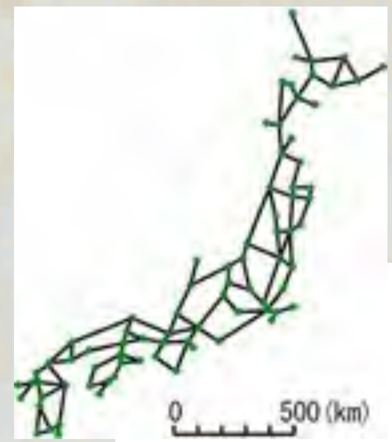
Mapping methods

Cartograms

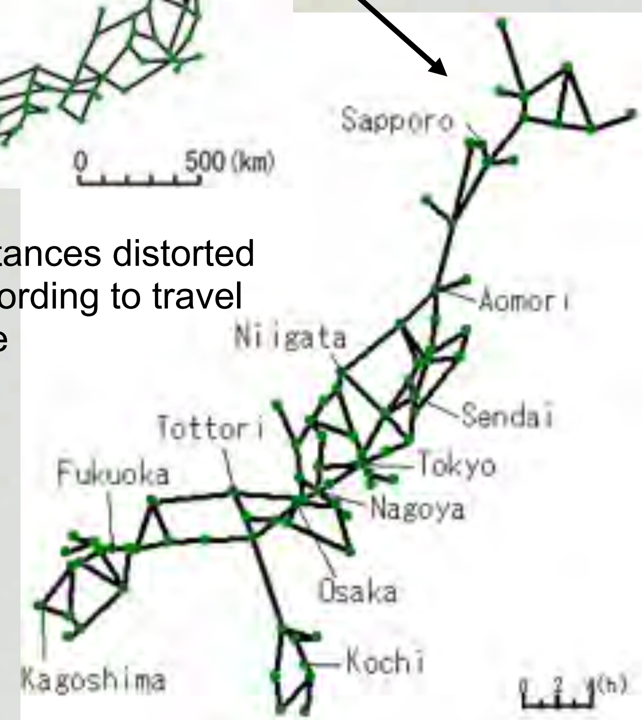
Not proper maps! The geographical positions of objects are distorted in order to better show the observed phenomenon.

A **cartogram** is a generalization of an ordinary thematic map, which is distorted by resizing its regions according to a geographically-related input parameter.





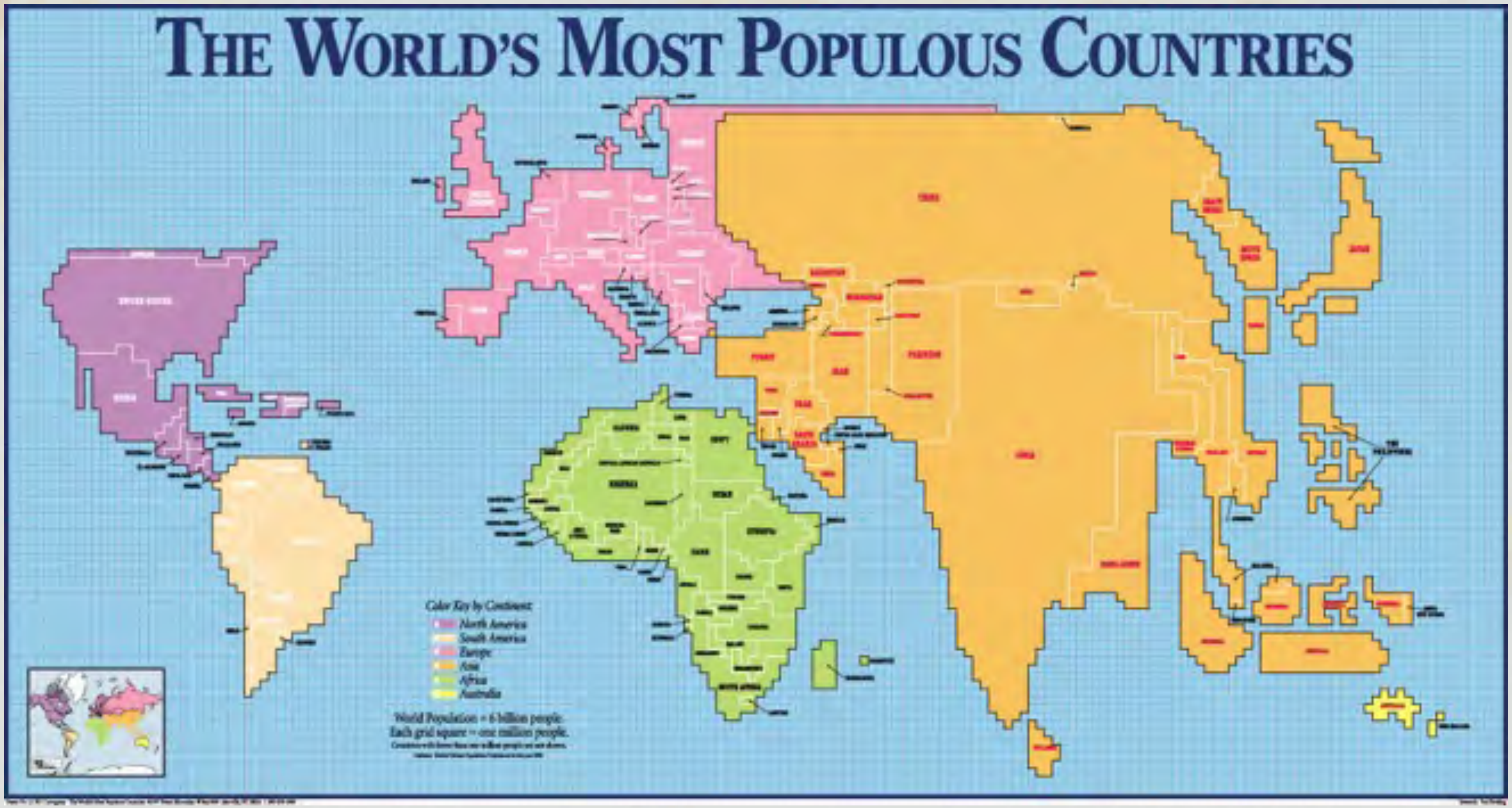
Geographically correct Japanese railway network



Distances distorted according to travel time

Mapping methods

Area cartogram: the area of each polygon represents some numerical attribute of the polygon (in this case the population).



Mapping methods

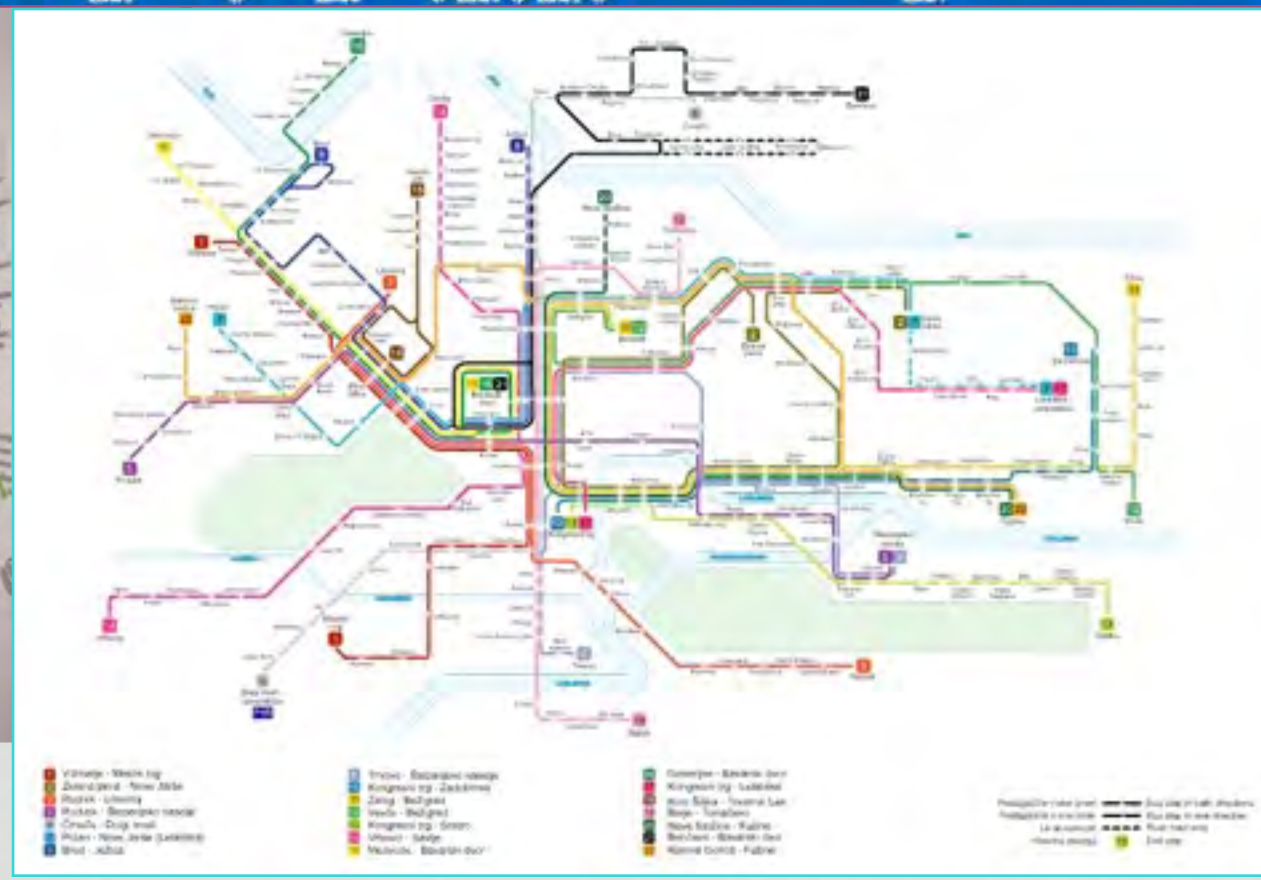
Distance cartogram: the shapes are resized according to some other alternative distance (non-geometric) from the source (in this case distance = price of air-fare from London)

Mapping methods

A linear cartogram: shows the location of the stations in relation to the public traffic train network in Stockholm

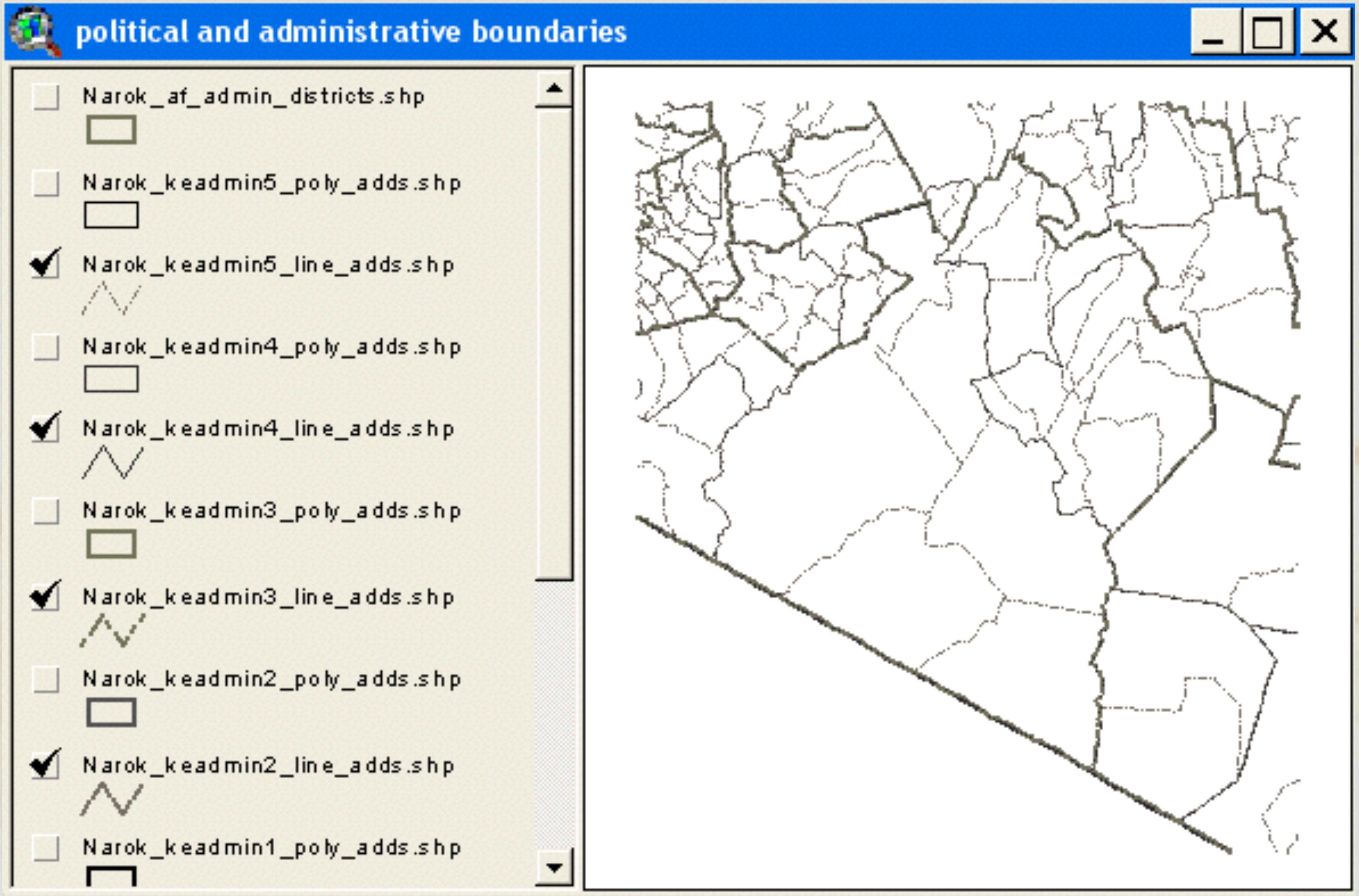
d1 = d2 < d3 Really????

Mapping methods

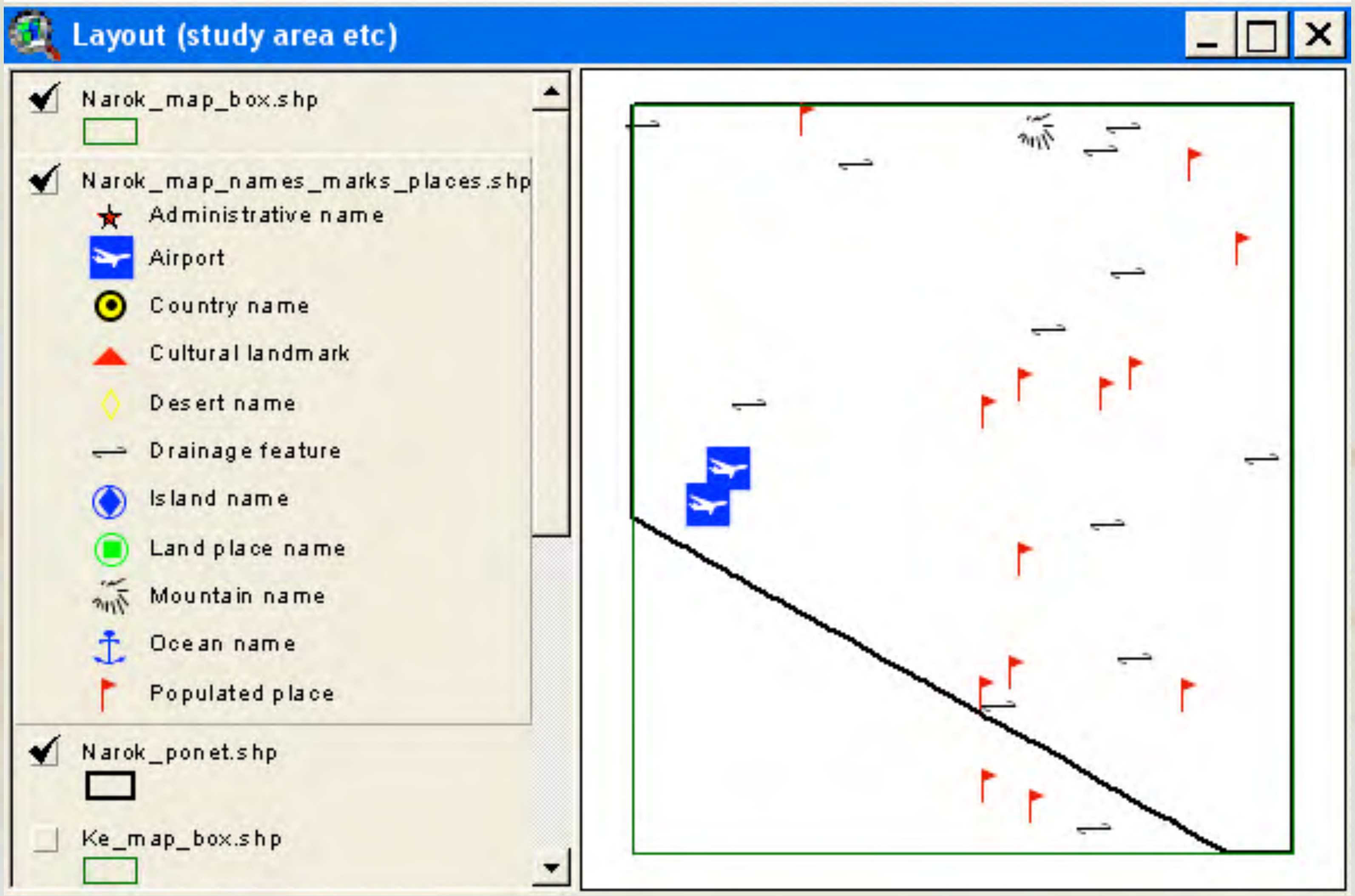


Linear cartograms are common as visualisations of public traffic networks.

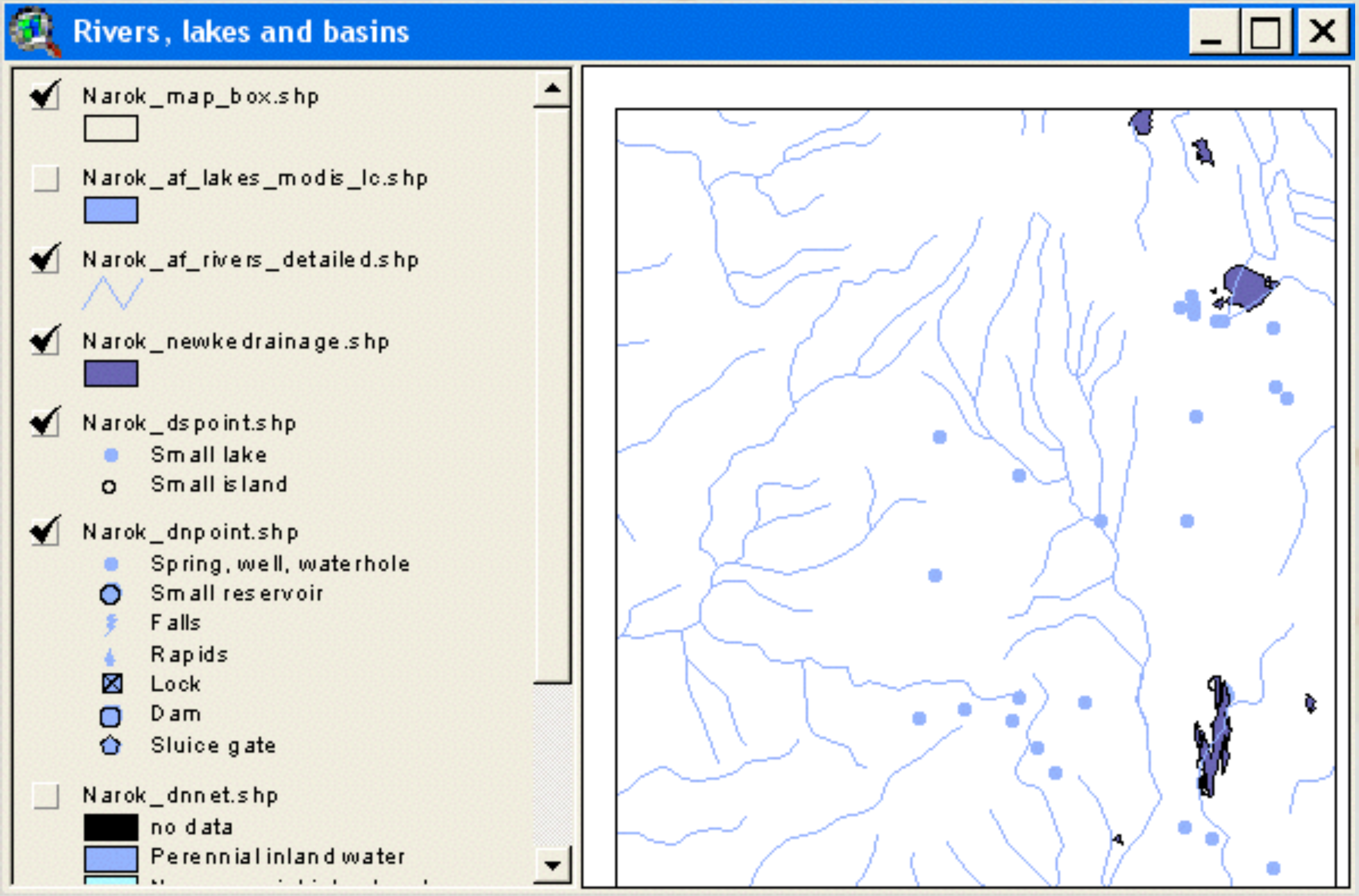
Example - Thematic mapping for vegetation analysis



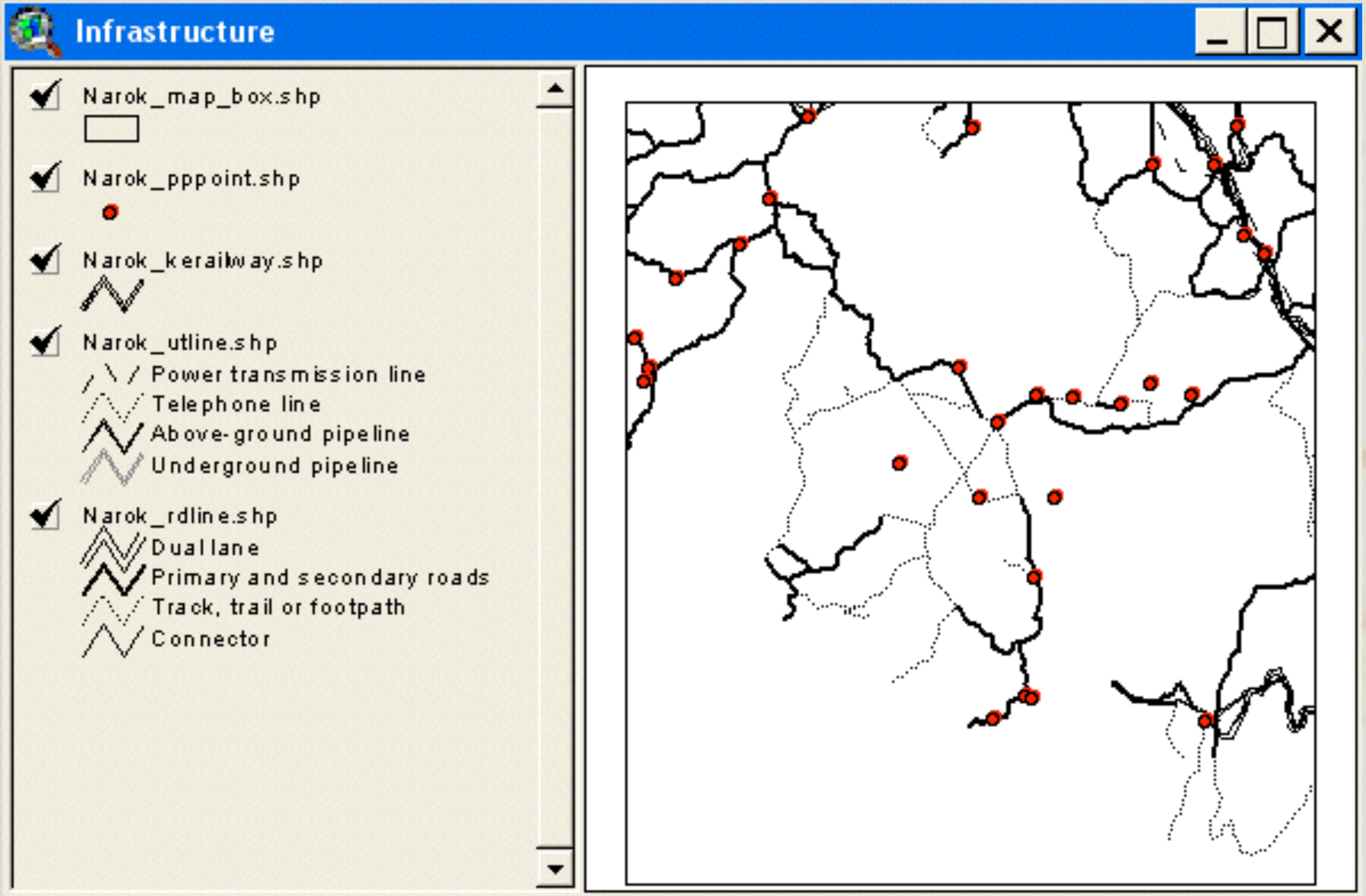
Example - Thematic mapping for vegetation analysis



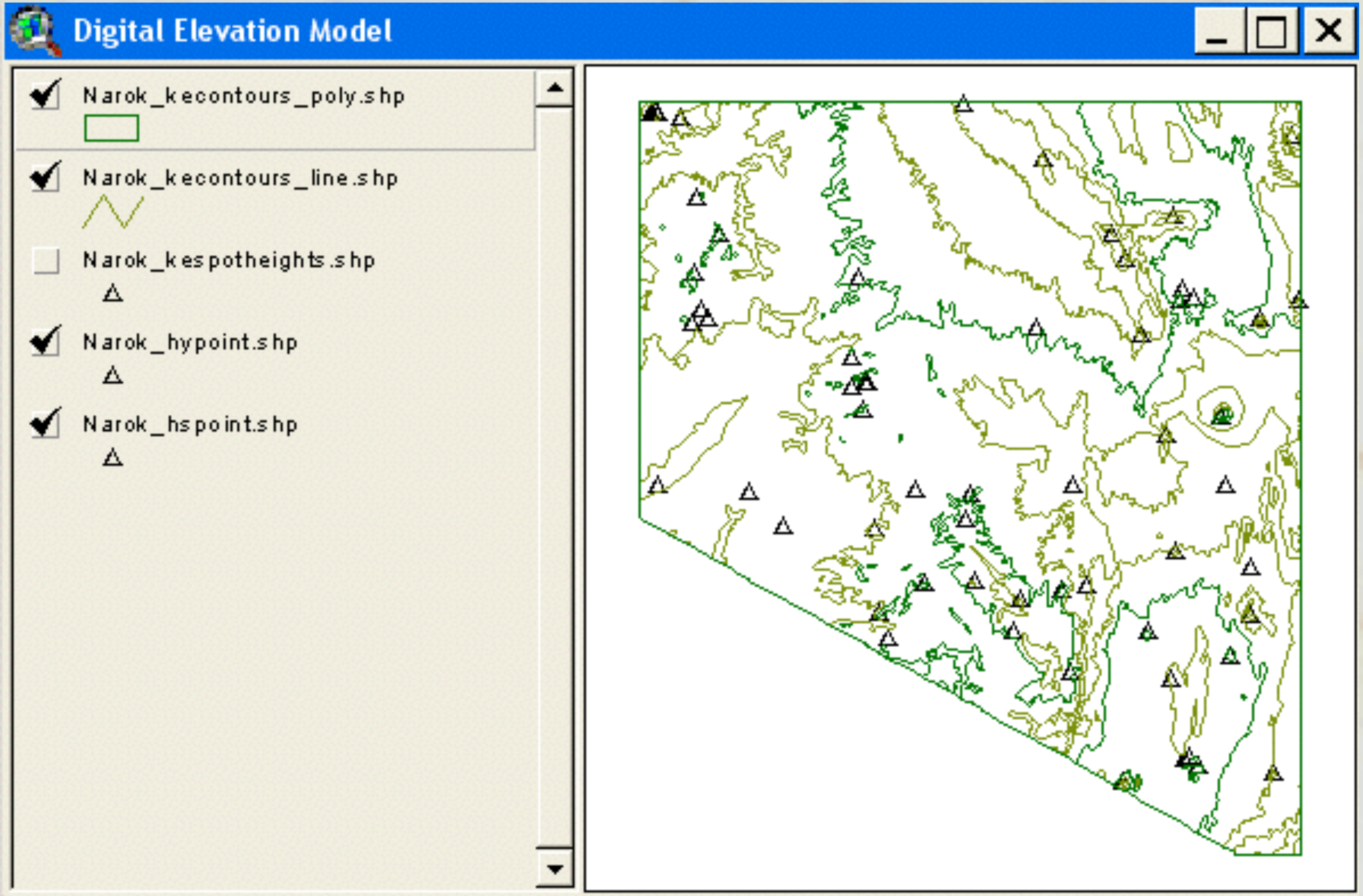
Example - Thematic mapping for vegetation analysis



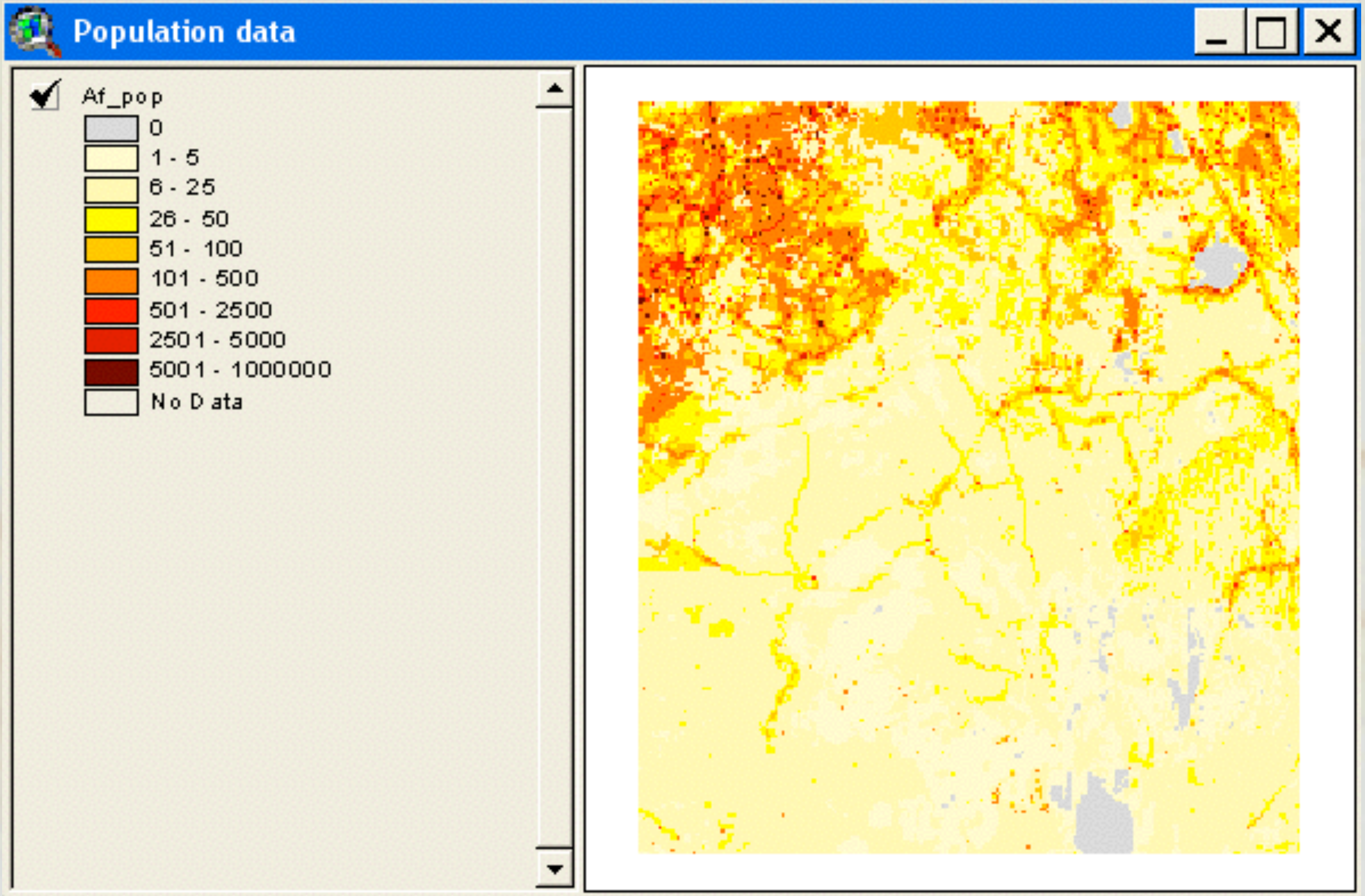
Example - Thematic mapping for vegetation analysis



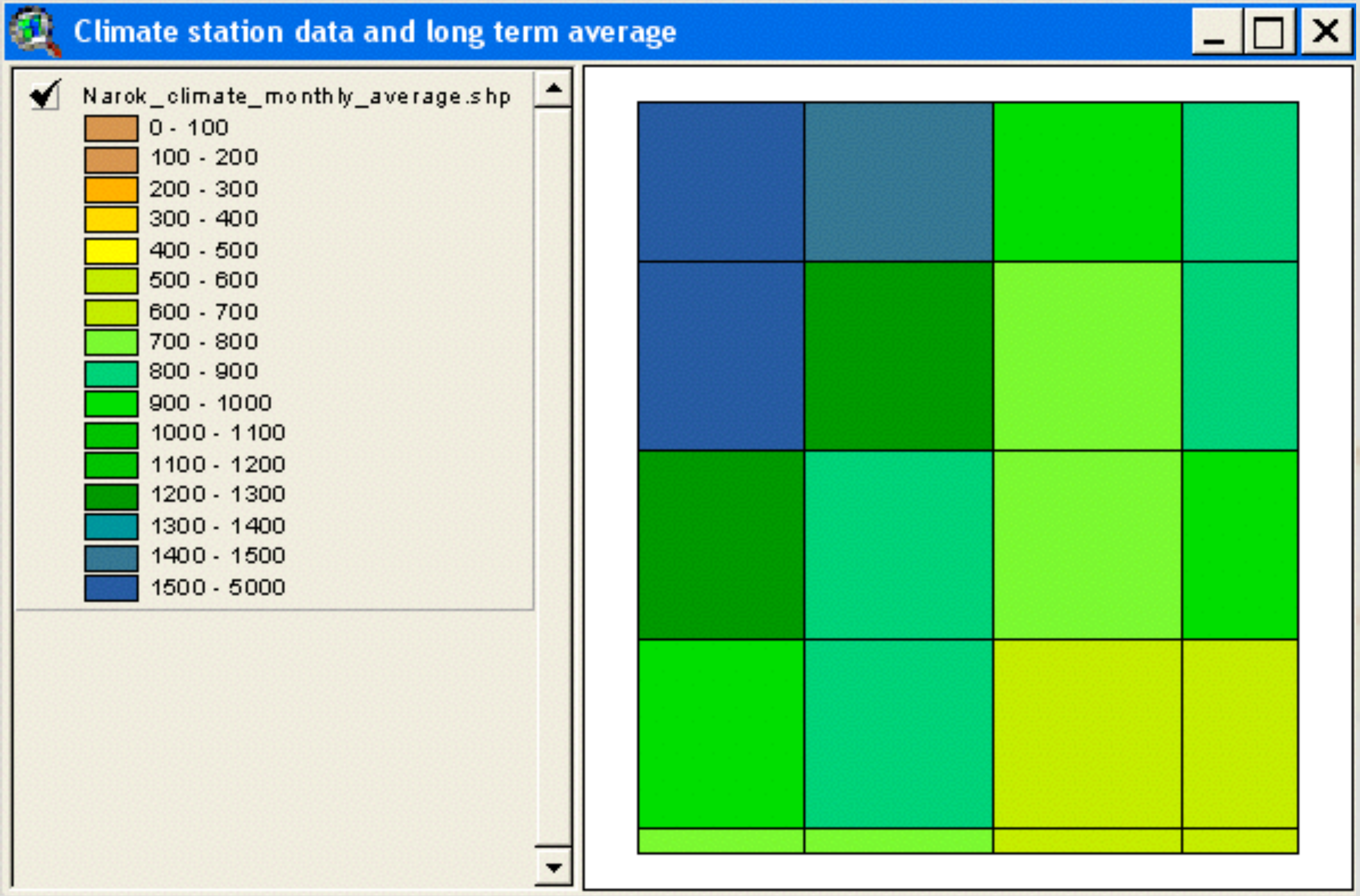
Example - Thematic mapping for vegetation analysis



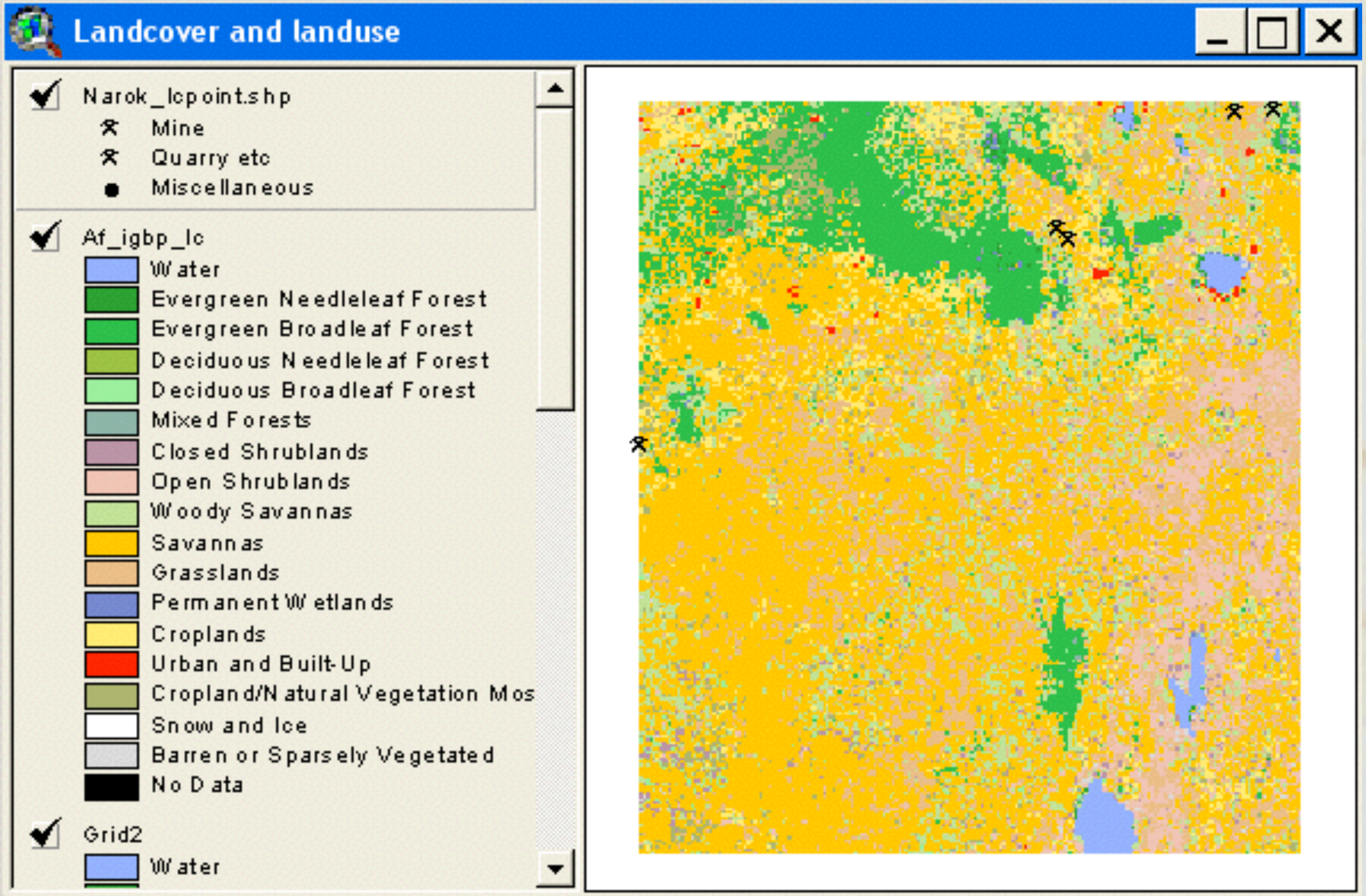
Example - Thematic mapping for vegetation analysis



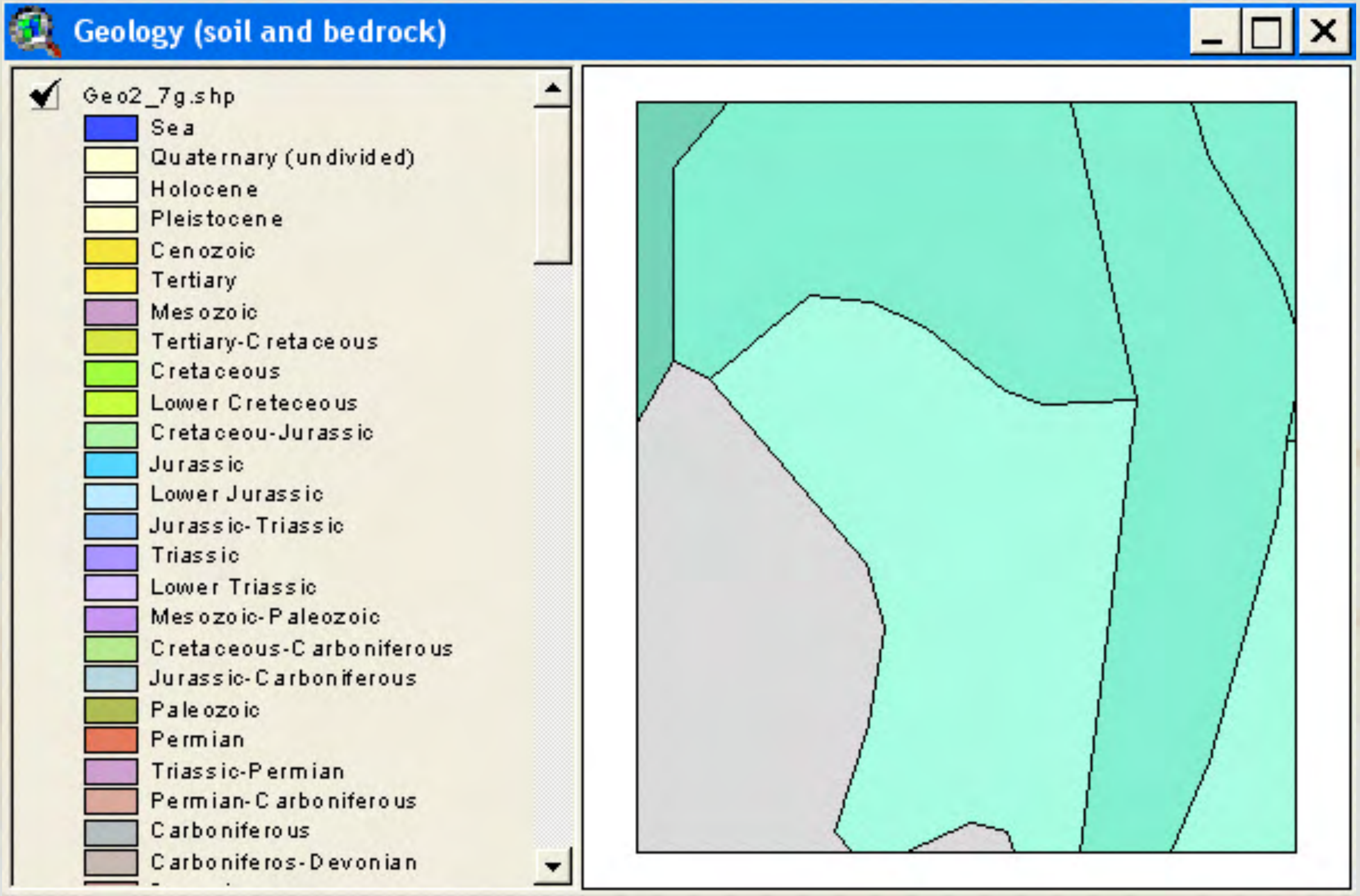
Example - Thematic mapping for vegetation analysis



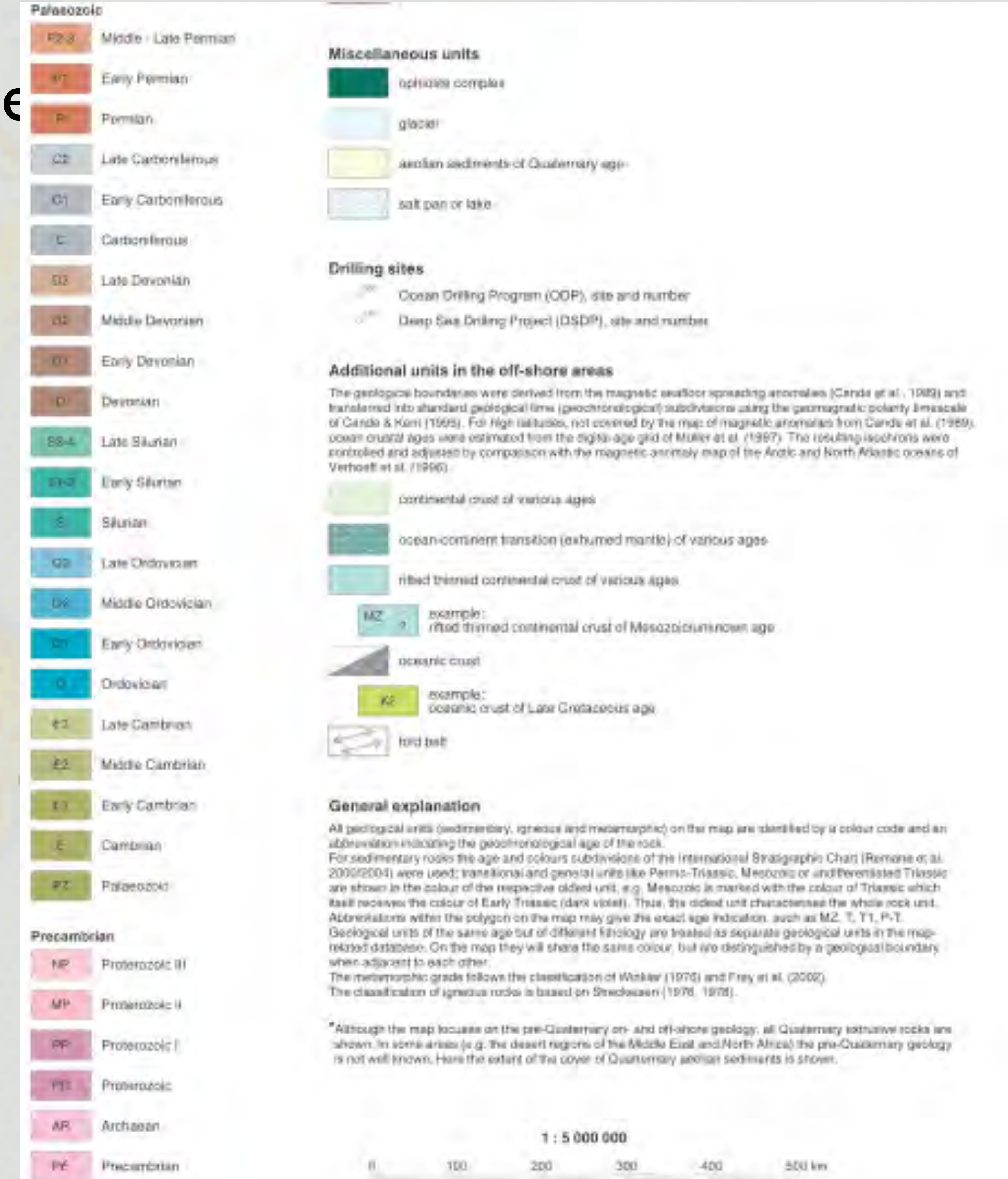
Example - Thematic mapping for vegetation analysis



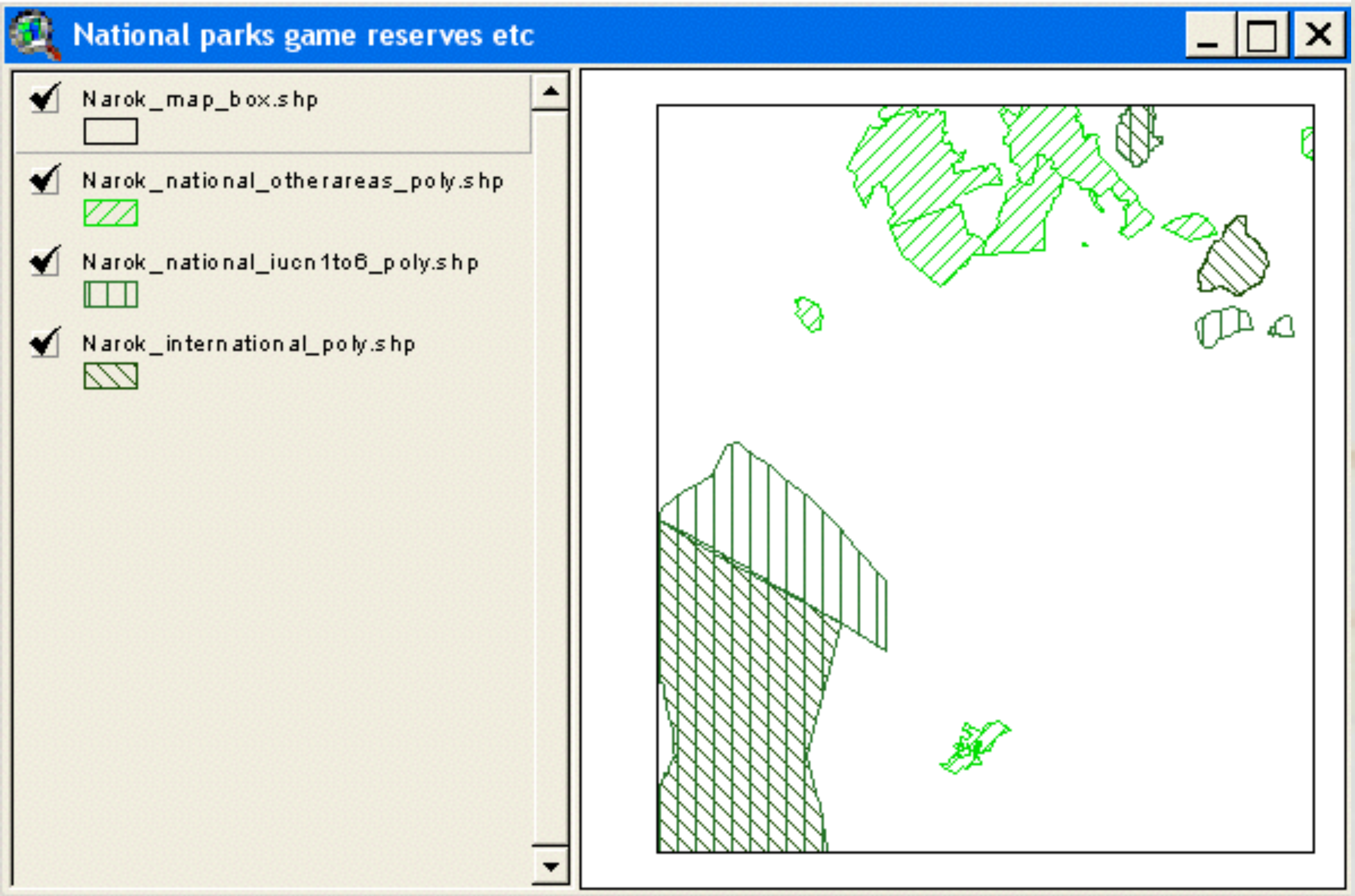
Example - Thematic mapping for vegetation analysis



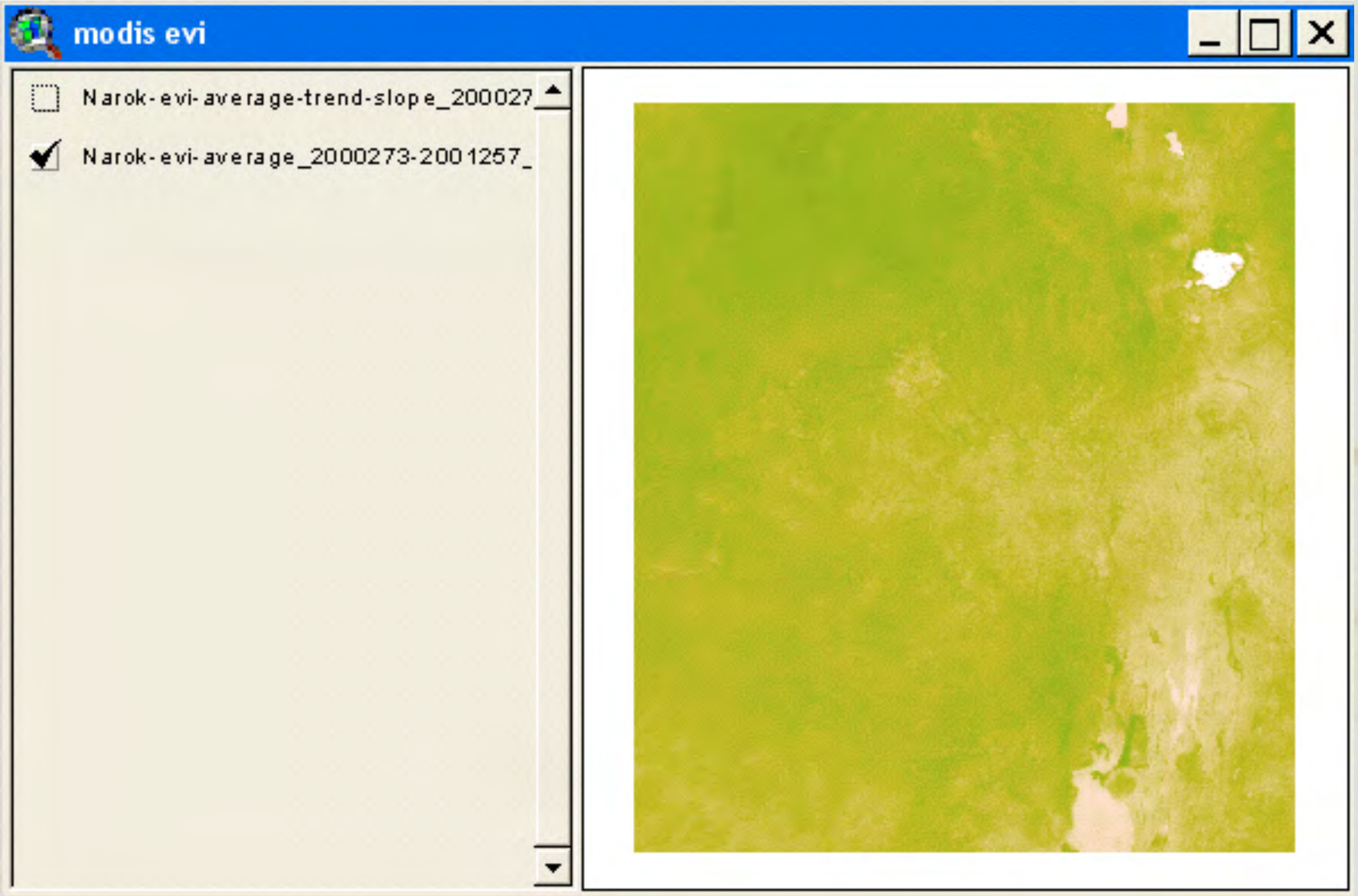
Example … analysis



Example - Thematic mapping for vegetation analysis



Example - Thematic mapping for vegetation analysis



Example - Thematic mapping for vegetation analysis