Geographic Information Technology Training Alliance (GITTA) presents:

Statistics for Thematic Cartography

Responsible persons: Boris Stern, Lorenz Hurni , Samuel Wiesmann, Yvonne Ysakowski

Content

1. Statistics for Thematic Cartography

Obtaining appropriate data is an important first step towards successful map communication. However, more important is the second stage: the processing you are going to perform on the data.
This learning unit will introduce you to various basic statistical techniques, which cartographers commonly use in mapping.

Learning Objectives

- Know why statistical applications are so essential in cartography.
- Be able to recognise the different levels of measurement.
- Know how and when to standardise data.
- Know how and when to classify data.

1.1. Basic Statistic Rules

Data level of Measured Phenomena

When measuring geographic phenomena in the field for the collected dataset, we commonly speak about the level of measurement. These out-coming data usually are divided into four levels of measurement:

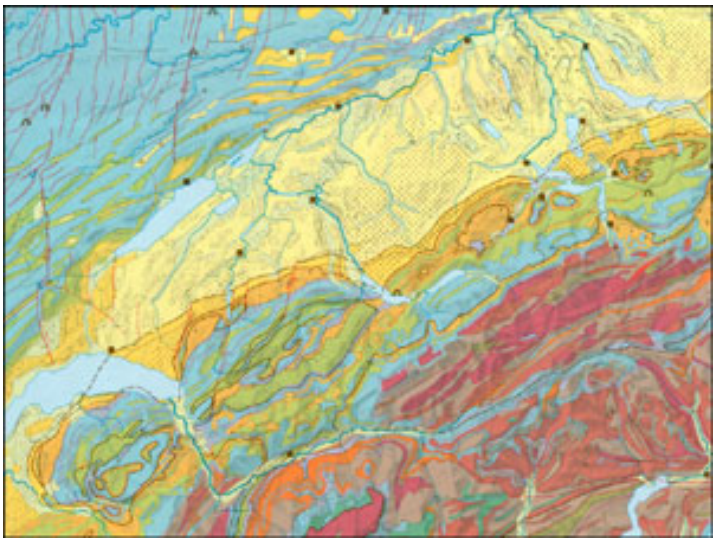
- Nominal data
- Ordinal data
- Numeric data
    - Interval data
    - Ratio data

The distinction of these data levels is relevant for the type of map presentation we wish to choose. Ordinal, interval, and ratio can be combined to create quantitative data.
In contrast, to these three data levels, nominal data only describes qualitative information. In the following paragraphs, we will point out the characteristics of each data level.
Thus, in this lesson, for which you need no prerequisites, you will learn the different levels of data measurement. In addition, you will also learn to prepare data for classification, as well as some basic classification rules.

1.1.1. Nominal Data

The main characteristic of nominal data levels is that the collected qualitative data can be grouped into different categories.
It is not possible however, to order or rank the data, as the data consists of value-free qualitative information.

Nominal data examples



Nominal Data Example (Spiess 1993)

Example: Geological maps are based on nominal data. Geological maps reveal the distribution of different geological patterns. We distinguish categories like granite, gneiss, limestone, tertiary and quaternary sediment deposits.

1.1.2. Ordinal data

Ordinal data is another measurement level. This data level includes both, the categorisation of data, and the

ordering (ranking) of data. The ranking of data, however, does neither rely on equal category intervals, or are the chosen categories indicated by numbers, but by a qualitative ranked class description.

Ordinal data example

Ordinal Data Example (Buri et al. 1999)

Example: Hazard map of snow avalanches. The three ranks stand for hazard classes: red for high danger, blue for moderate danger, and yellow for low danger.

1.1.3. Numeric data

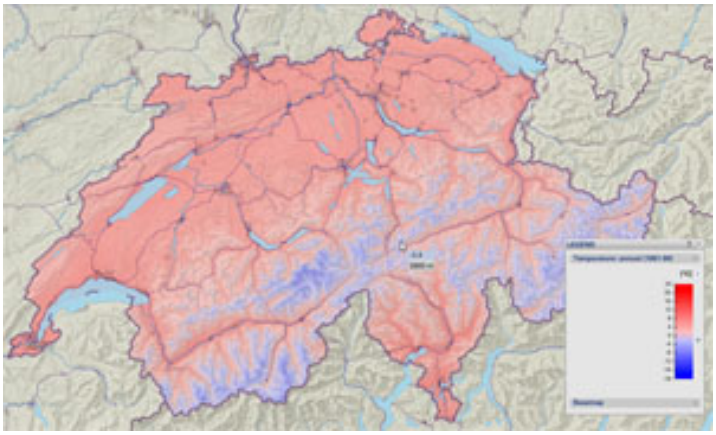
Contrary to nominal and ordinal maps, the numeric map reveals the quantitative character of collected data. Regarding numeric map information, we have to consider two different types of data: Interval and Ratio data.

Interval data

The interval level of collected data have three characteristics:

- The categorisation of data
- The ordering of data, but with...
- ...explicit numerical indication of the categories value

Interval example



Interval Data Example, Atlas of Switzerland 3 (Bundesamt für Landestopografie swisstopo (eds.) 2010)

Example: Temperature in Switzerland (mean annual temperature 1961-1990, degree celsius). A classic example for interval data is the temperature scales of CELSIUS and FAHRENHEIT. Both scales consist of ordered values and reveal the precise difference between their temperature values. However, the problem of these interval scales is the arbitrary character of their zero points.

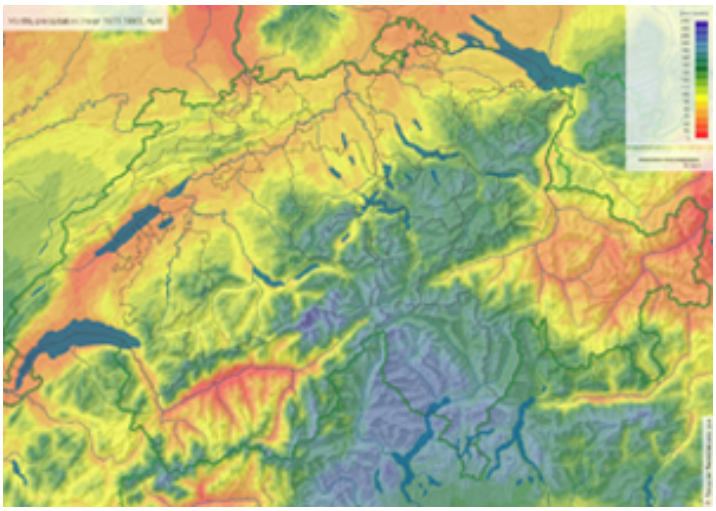
Ratio data

Ratio data has the same characteristics as interval data:

- The categorisation of data
- The ordering of data, but with
- The explicit numerical indication of the value differences between the categories.

However, contrary to interval data, ratio data has no arbitrary but an absolute zero point. In many fields, it is very common to work with ratio data.
Continuing with the example of temperatures, we state that the KELVIN scale has an absolute zero point and therefore has a ratio scale. This absolute zero point of temperature is the lowest possible temperature of anything considered in the universe. Thus, we are allowed to say that 30°K is twice as warm as 15°K, also in terms of kinetic energy of molecules (Slocum 1999).

Ratio example



Ratio Example, Atlas of Switzerland 1 (Bundesamt für Landestopografie swisstopo (eds.) 2000)

Example: Precipitation in Switzerland (average April, in mm). The registration of precipitation has an absolute zero point too. If there is no rain falling, we measure 0 mm of precipitation. So we can say that 1000 mm is twice as much as 500 mm because precipitation relies on ratio data.

1.1.4. The Importance of Classification

Classification allows you to structure the thematic communication message. How well this is done depends largely on your ability to understand the geographic phenomenon. The classification operation behaves pretty much like a group of stacked sieves. Each sieve acts as a class boundary, and only values of certain sizes are allowed to pass into one of several classes (Dent 1999).

The sieve analogy in classification.
Each sieve functions as a screen allowing only balls of a certain size to drop through to the next level. Each sieve can be compared to a taxonomic criteria established for the particular study. Spaces between the sieves become the classes, and the sieves are the class boundaries (Dent 1999).

Discover interactively the sieve analogy
Discover interactively the sieve analogy: click on "Classify Data" to start classification.

Only pictures can be viewed in the PDF version! For Flash etc. see online version. Only screenshots of animations will be displayed. [link]

1.1.5. Data Preparation

What is the representation purpose?
Before beginning any work on the data, it is important that you know exactly what is the information you want to convey, to whom and why?
It is advisable in particular to deepen the knowledge of the readers for whom the map is intended, by defining at the same moment their "reading potential" and their interest for the treated subject.

Choice and manipulation of the data
Before proceeding to the cartographic representation of the data, it is necessary to choose which data to represent, and then to treat the data to make them capable of being represented on a map.

Data Preparation
Data preparation includes the following stages:

- Arrange the values into increasing order,
- Calculate some data information, such as:
  - The number of data (N)
  - The extreme values
  - Indications of descriptive statistics, such as mode, median, and average
  - Dispersal parameters: mean, co-variance and variance
- Complete this statistical information with graphs, such as
  - Distribution diagram
  - Histograms

Nowadays, various software (MS Excel for example), allow easy and fast data preparation.

Notice: the distribution diagram is the very basic data preparation you have to do. It will help you to choose the adequate classification method.

1.1.6. Basic Classification Rules

How to Define the Number of Classes
The number of classes is to be connected to the objectives assigned to the cartographic representation.
If the purpose of the map is to show simply progress, a continuous pressure gradient (as for example the increase of the height on a topographic map) then the number of classes can be relatively important. On the opposite, if it is to put classes in an obvious place for having a meaning for the author, then it is indispensable that the consequential classes and the graphic symbols have a correspondence without ambiguity. In this case, the reader has to have the possibility of differentiating well beaches. Then the number from 5 to 6 appears as a maximum.

How many classes per map?
Generally, it is difficult to visually perceive information, which divides up into several classes or categories. This is why the number of classes must be big enough to be useful, but limited to allow the map to be readable without any trouble.
Generally, a distribution between 4 and 8 classes is correct and 5-6 classes is the standard.

How to Define the Classes

There are several ways to define classes. You should still choose the method which suits best and for the data distribution and for the aspect of the data you want to emphasise. In any case, with the method you choose:

- The classes have to cover all the data values in order that the limits are contiguous to the data.
- All classes should be homogeneous.

Further Classification Rules

- A value has to appear in a class and in only one class. The class limits must not overlap.
- You should not present the values of the limits of classes with a significant number of decimals superior to the one that allows the exactness of the data.

1.1.7. Summary

The distinction of data levels is relevant for the type of map presentation you wish to choose. Ordinal, interval, and ratio can be combined to create quantitative data.

1.2. Standardisation and Classification

A common problem faced by map-making is whether raw data should be classed into groups, and if so, which method of classification should be used.
In this learning unit, you will learn the different kinds of data standardisation, why to use classification, and the different kinds of classifications.

You need to go through the Basic Statistic Rules unit, before proceeding further with this unit.

1.2.1. Standardisation of Data

The standardisation of data is necessary for more meaningful thematic map presentation (e.g. costs per inhabitants instead of the total of absolute costs). The resulting common denominator enables a comparison between different types of collected data.
Moreover, we standardise data if we wish to show the relation between our collected data and another dataset. Thus, we standardise data in order to make our data comparable to others, to show the ratio, and enable a better analysis of our data.
In the next paragraph, we will introduce different standardisation approaches for numerical data.

Ratio standardisation
The most common method of standardisation approaches is the simple ratio standardisation. For ratio standardisation, we divide an area-based numerical dataset by another area-based numerical dataset. As the numerator and the denominator both consist of the same measurement units, the result is a proportion. This proportion can also be expressed as a percentage. An example may be the division of the water run-off of a considered catchment area [mm/m2] by the measured precipitation of the same catchment area [mm/m2]. The resulting ratio between run-off and precipitation is the "run-off coefficient" of the catchment area [%].

Density standardisation
When standardising data to indicate the density, we divide a non area-based variable by an area. For example, If we want to calculate the soil erosion per area [kg/m2], we have to divide the measured eroded soil weight [kg] by the area size [m2].

Rate standardisation
Another way of standardising data is to compute the ratio of two non area-base variables. Accordingly, the resulting units are always rates. Example: We may calculate the damage costs per person after a natural hazard [USD / person]. Another example of rate standardisation is the calculation of the available hospital beds per inhabitant [hospital bed/inhabitant].

Area-based rate standardisation
Sometimes we standardise data by dividing an area size by a non area-based variable. Example: When we divide the total size of a village's estates [m2] by the number of land owners [persons], the emerging result is the average size of estate per land owner [m2/landowner], which is a so-called area-based rate (Slocum 1999).

Additional remark

It is sometimes necessary to standardise data before we can compare it to other datasets. However, we do not need to standardise our data, if we compare our collected data with data that refers to one and the same basic measurement settings (e.g. same location, same measurement method,

etc.).

Example: If we measure the precipitation at a specific place x over one year each day, we can compare these measurements without standardising them. We can even use an invented unit for comparison, as long as the measurement system and its units remain the same during the considered year. However, as soon as we want to compare our measured precipitation data with the precipitation data of another place, we need to standardise the data in order to obtain the same units (a common denominator).

1.2.2. Classification of Data

Classified versus unclassified data

PLEASE NOTE: this learning object "Classification of Data" is currently under revision. Several sections are therefore missing at the moment (marked with **[...]**). The remaining text is mainly based on (Slocum 1999).

As you have seen in the previous paragraph about "data levels", numerical data consists of the exact indication of measured information. As you may imagine, such measurable information is very important for geographical data analysis and for precise value presentations on maps. However, for an optimal analysis of numeric data we sometimes need to classify our dataset with a method for an appropriated thematic map presentation, which allows an optimal map analysis of numeric data. In this section, we will reveal when we need to classify data, and when we can work with unclassified data.

What is a classified map?

A classified map represents data that has been grouped into different classes. On the map, the different classes can be distinguished e.g. by different colours (hue, brightness, or saturation).

Why can it be useful to classify data before creating a map?

The human eye only has a limited ability to discriminate a large number of different areal symbol shades. Due to this fact, it is sometimes essential to classify quantitative thematic map content. This allows us to create a smaller number of data classes and to choose symbol shades that can be distinguished easily.

What is the difference between a classified and an unclassified map?

Classified maps consist of colour shades that are generally based on the conventional "maximum-contrast" approach, using equally spaced tones from one class to another. Thanks to this method, the classified map does not reveal a huge and inhomogeneous range of colour variations.

[...]

Thus, we finally have to decide when we choose to classify our collected data and when not. You should have considered two criteria when you decide whether you create a classified or an unclassified map presentation:

1. Do you wish to maintain numerical data relations? If you wish to create a map that maintains the data relation, unclassified data theoretically does a better job than classified data, as unclassified data allows us to maintain the numerical relations between data. This means that the colour shades on an unclassified map are directly proportional to the values of each enumeration unit.
2. Is your map intended to be used for data presentation, or is it meant to be applied for data analysis? When you create a map that will be used for a simple data presentation, you basically have the choice between either classified or unclassified data. When you decide to use classified data for your map, however, the differences in value and colour shade usually become more obvious. In general, cartographers do not approve the result of non-classified data, since unstructured (or non-generalised)

maps are composed of many individual symbols. Quite contrary to this, it is advisable not to use more than six classes, so that the map reader is able to distinguish them easily.

However, if the map you create is intended for data analysis, it is worth comparing a large variety of visual classification approaches. You do this in order to choose the best method for your specific thematic analysis. This map comparison may possibly include unclassified maps, too.

Major Classification Methods

For thematic map presentation, the acquired and analysed thematic data values are often grouped into classes, which simplify the reading of the map as we have learned in the previous section. If you decide to classify your data, you may wonder, what would be the best method. For this purpose, we will repeat and refresh the basics of your knowledge about statistical methods in the following. The major methods of data classification are:

- Equal intervals,
- Mean-standard deviation,
- Quantiles,
- Maximum breaks and
- Natural breaks

[...]

The Equal Interval Classification (constant class intervals)

In this classification method, each class consists of an equal data interval along the dispersion graph shown in the figure. To determine the class interval, you divide the whole range of all your data (highest data value minus lowest data value) by the number of classes you have decided to generate.

After you have done that, you add the resulting class interval to the lowest value of your data-set, which gives you the first class interval. Add this interval as many times as necessary in order to reveal the number of your predefined classes.

When is it useful to choose the method of equal class intervals?

It is appropriate to use equal class intervals when the data distribution has a rectangular shape in the histogram. This, however, occurs very rarely in the context of geographic phenomena. Moreover, it is useful to use this method when your classification steps are nearly equal in size. The major disadvantage of this method is that class limits fail to reveal the distribution of the data along the number line. There may be classes that remain blank, which of course is not particularly meaningful on a map.

[...]

The Mean-Standard Deviation Classification

Another method that allows us to classify our dataset is the standard deviation. This method takes into account how data is distributed along the dispersion graph. To apply this method, we repeatedly add (or subtract) the calculated standard deviation from the statistical mean of our dataset. The resulting classes reveal the frequency of elements in each class.

The mean-standard deviation method is particularly useful when our purpose is to show the deviation from the mean of our data array. This classification method, however, should only be used for data-sets that show an approximately "standardised normal distribution" ("Gaussian distribution"). This constraint is the major disadvantage of this method.

[...]

The Quantiles Classification
Another possibility to classify our dataset is to use the method of quantiles. To apply this method we have to predefine how many classes we wish to use. Then we rank and order our data classes by placing an equal number of observations into each class. The number of observations in each class is computed by the formula:

If no integer values are resulting from this division, we attempt to place approximately the same number of observations in each class.
An advantage of quantiles is that classes are easy to compute, and that each class is approximately equally represented on the final map. Moreover, quantiles are very useful for ordinal data, since the class assignment of quantiles is based on ranked data. The main disadvantage of this classification method are the gaps that may occur between the observations. These gaps sometimes lead to an over-weighting of some single detached observations at the edge of the number line.
[...]

The Maximum Breaks Classification
When we choose to use the method of maximum breaks we first order our raw data from low to high. Then we calculate the differences between each neighboring value, when the largest value differences will be applied as class breaks. You can also recognise the maximum breaks visually on the dispersion graph: large value differences are represented by blank spaces.
One advantage of working with this method is its clear consideration of data distribution along the number line. Another advantage is that maximum breaks can be calculated easily by subtracting the next lower neighboring value from each value. A disadvantage, however, is that the systematic classification of data misses any proper attention to a visually more logical and more convenient clustering (see "Natural breaks").
[...]

The Natural Breaks Classification
Applying the classification method of "natural breaks", we consider visually logical and subjective aspects to grouping our data set. One important purpose of natural breaks is to minimise value differences between data within the same class. Another purpose is to emphasize the differences between the created classes.
A disadvantage of this method is that class limits may vary from one map-maker to another due to the author's subjective class definition (Slocum 1999). The Jenks-Caspall Alorithm formalizes this procedure and is often used in GIS software.
[...]

Discussion of the Classification Methods

Equal intervals
Particularly useful when the dispersion graph has a rectangular shape (rare in geographic phenomena) and when enumeration units are nearly equal in size. In such cases, orderly maps are produced.

Mean-standard Deviation
Should be used only when the dispersion graph approximates a normal distribution. The classes formed, yield information about frequencies in each class. Particularly useful when the purpose is to show deviation from the array mean. Understood by many readers.

Quantiles
Good method of assuring an equal number of observations in each class. Can be misleading if the enumeration units vary greatly in size.

Maximum Breaks
Simplistic method which consider how data are distributed along the dispersion graph and group those that are similar to one another (or, avoid grouping values that are dissimilar). Relatively easy to compute, simply involving subtracting adjacent values.

Natural Breaks
Good graphic way of determining natural group of similar values by searching for significant depressions in frequency distribution. Minor troughs can be misleading and may yield poorly defined class boundaries.

1.2.3. Test your knowledge about Standardisation and Classification

Test your knowledge about the Classification Rules for Diagrams and Charts with the following Self-Evaluation:

Only pictures can be viewed in the PDF version! For Flash etc. see online version. Only screenshots of animations will be displayed. [link]

1.2.4. Summary

In this learning unit, several data classification methods and criteria for selecting an adequate method were examined. These methods that consider how data are distributed along the dispersion graph (such as Natural Breaks and Optimal) are preferable because they place similar data values in the same class (rather than similar data values in different classes).
Methods that do not consider the distribution of data along the dispersion graph (such as Equal Intervals and Quantiles) may however also satisfy other criteria (Slocum 1999).

1.3. Statistics for Thematic Cartography Evaluation

Your knowledge about "Statistics for Thematic Cartography" will be evaluated here. Using the map and the density data, you should create a map of Europe's population densities.
You should consider the following rules:

- Choose the best classification and explain why you have chosen it (2-3 lines of text).
- Create a digital map, which considers the main cartographic elements and steps: layout, colours, etc.
- Send the results to your tutor.

You can download the data here:

- Density of population in Europe (1993): Density_Pop_Europe.txt(2 Ko)
- Basemap of Europe:
  - europe.ai(243 Ko)
  - europe.svg(183 Ko)
  - europe.wmf(96 Ko)

1.4. Summary

In this lesson, you have learned about the data level of measured phenomena, the common methods of data standardisation and classification of data and how and when to use them. As well as the various classification rules for diagrams and charts. Thus, you should be able to handle statistics for thematic cartography on your own now.

1.5. Recommended Reading

- **DENT, B. D.**, 1999. *Cartography - Thematic Map Design*. 1st Edition. WCB McGraw-Hill. [electonically reproduced by permission of Pearson Education, Inc., Upper Saddle River, New Jersey]
  Chapter "Techniques of quantitative thematic mapping" especially
- **LEWIS, P.**, 1977. *Maps and Statistics*. Methuen & Co Ltd.
- **SLOCUM, T. A.**, 1999. *Thematic cartography and Visualization*. Prentice Hall Inc..
  Chapter "Data Classification" especially

1.6. Bibliography

- **BUNDESAMT FÜR LANDESTOPOGRAFIE SWISSTOPO (EDS.)** (2000). *Atlas of Switzerland interactive*. [CD-ROM]. First Edition.
- **BUNDESAMT FÜR LANDESTOPOGRAFIE SWISSTOPO (EDS.)** (2010). *Atlas of Switzerland 3*. [CD-ROM]. Third Edition.
- **BURI, H., KIENHOLZ, H, LINDER, E., ROTH, H., SCHWEIZER, M.**, 1999. *Achtung Naturgefahr, Verantwortung des Kantons und der Gemeinden im Umgang mit Naturgefahren*. Bern: Amt für Wald, Tiefbauamt, Amt für Gemeinden und Raumordnung.
- **DENT, B. D.**, 1999. *Cartography - Thematic Map Design*. 1st Edition. WCB McGraw-Hill. [electonically reproduced by permission of Pearson Education, Inc., Upper Saddle River, New Jersey]
- **LEWIS, P.**, 1977. *Maps and Statistics*. Methuen & Co Ltd.
- **SLOCUM, T. A.**, 1999. *Thematic cartography and Visualization*. Prentice Hall Inc..
- **SPIESS, E.**, 1993. *Schweizer Weltatlas*. Konferenz der Kantonalen Erzihungsdirektoren.