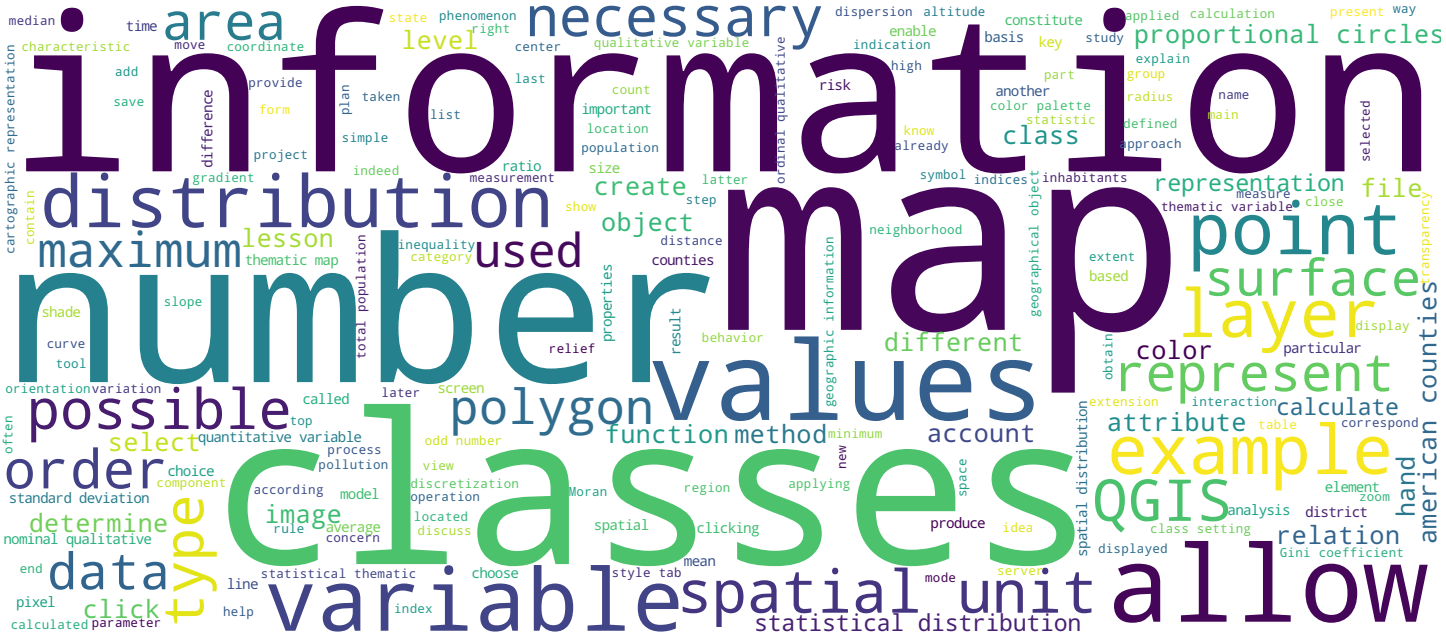
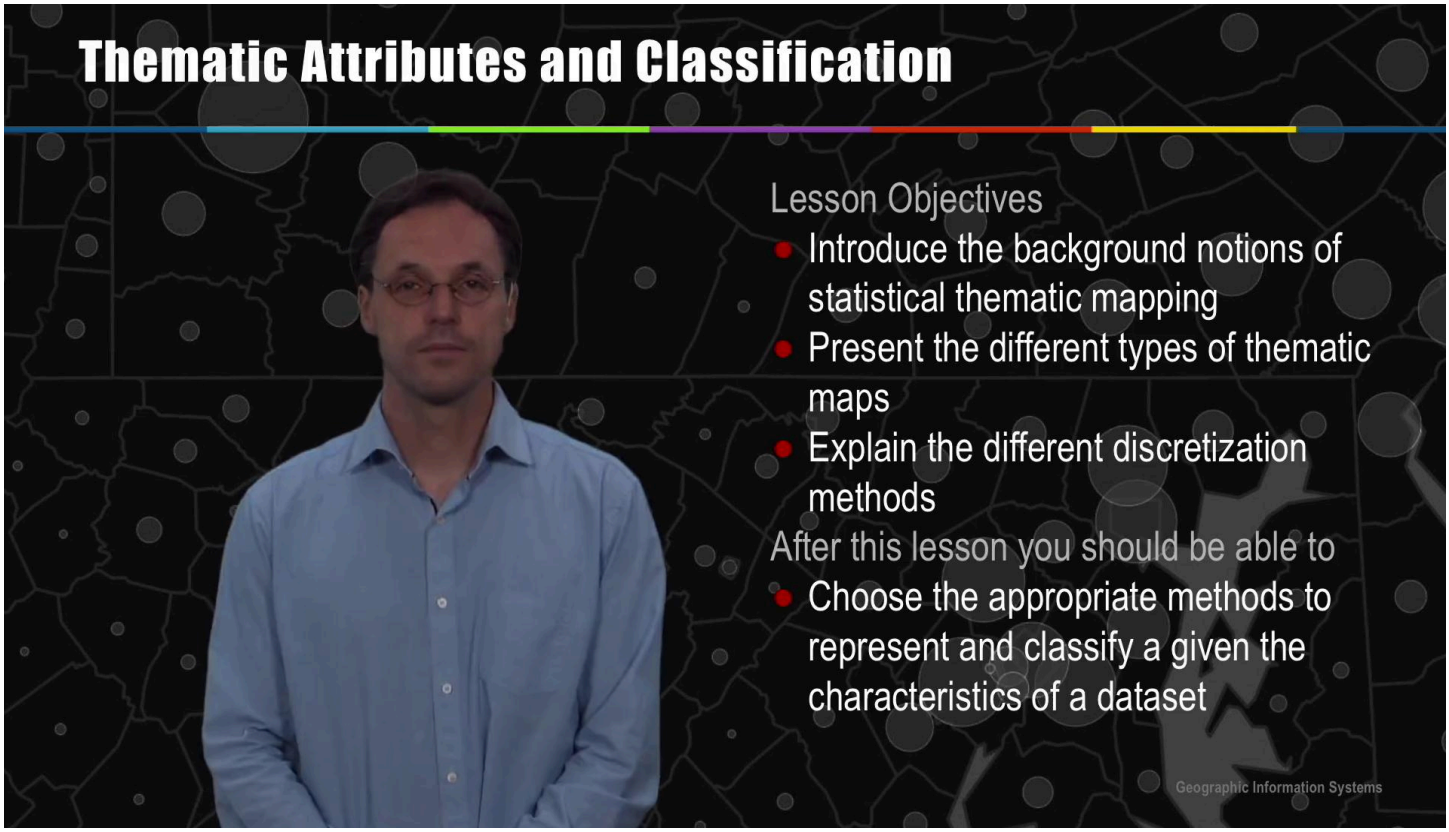
Thematic Mapping and the Semiology of Graphics
Thematic Attributes and Classification

An Introduction to Geographic Information Systems

Stéphane Joost, Marc Soutter, Fernand Kouamé, Amadou Sall



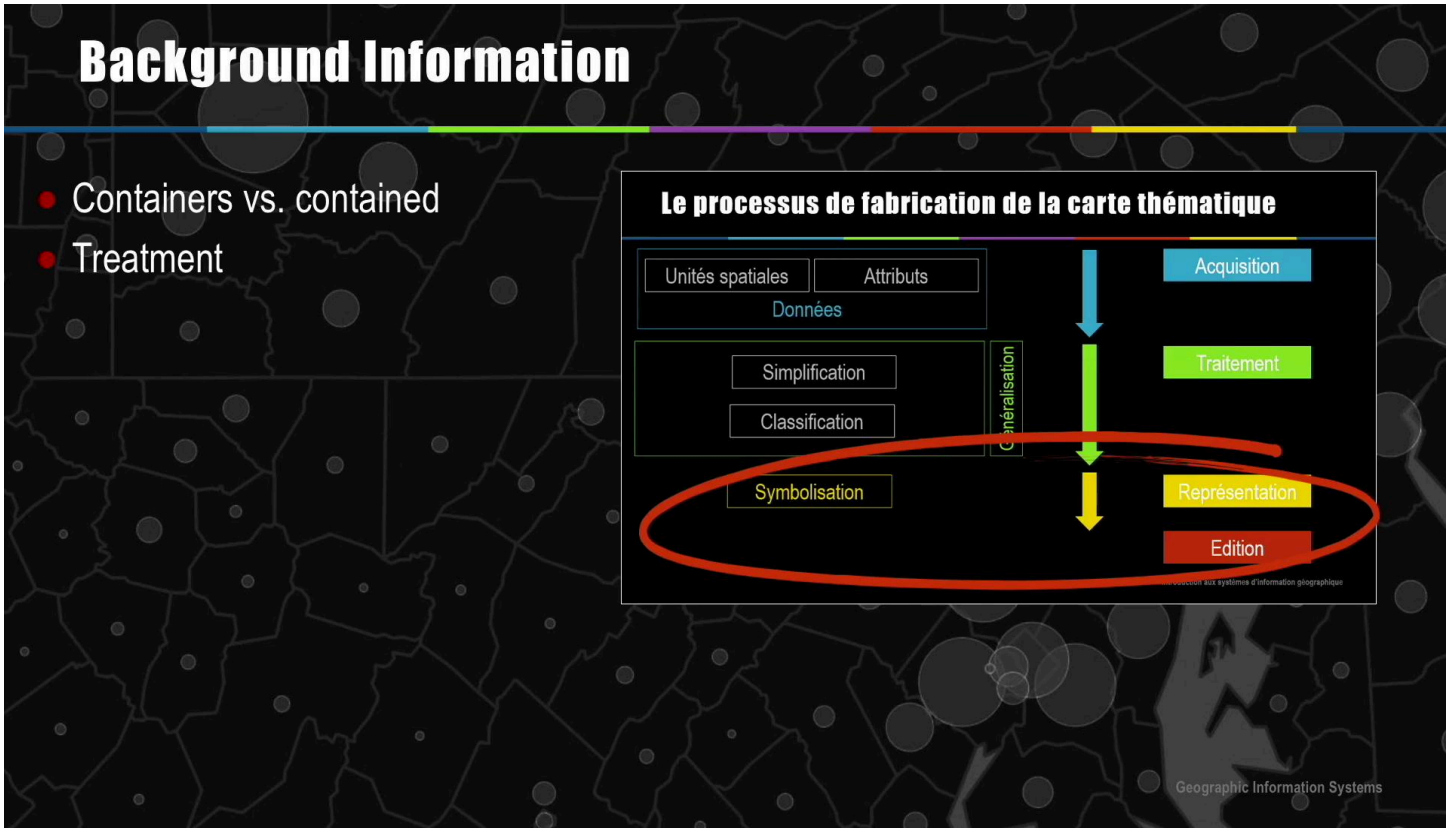


Good morning and welcome to this third lesson dedicated to thematic mapping and graphical semiology. We will discuss this time the practical aspects of the development of statistical thematic maps and examine how to take into account the different types of attributes of geographical objects to be represented. The goals of this lesson are to instill the necessary notions to the implementation of statistical thematic mapping, to present the different types of thematic maps and explain how the main statistical information classifying methods work. At the end of this lesson, you will be able to select the appropriate classifying and graphical representation methods depending on the characteristics of the data sets and build correct statistical thematic maps.

Notes

Summary

0m 30s

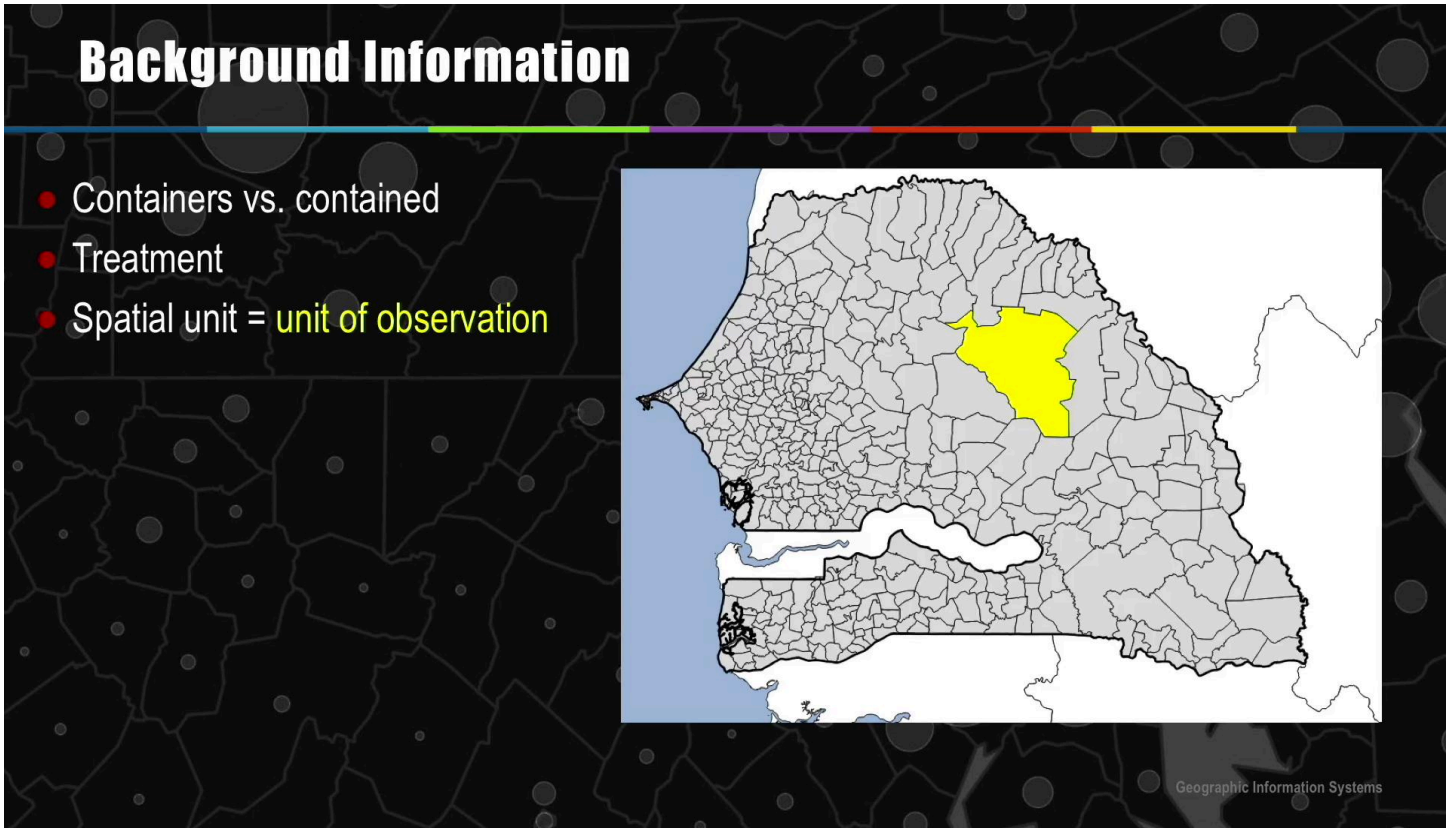


So far, we have mainly discussed the signs used to represent the geographic information, that is to say the container materialized by the coordinates X and Y. We will now address the notions necessary for the processing of the content, so the thematic component represented by the characteristics or attributes of geographic objects. You remember that the treatment phase contributes to the generalization of the information and that it contains a step of simplification and classification, which then makes the representation of the model of reality that constitutes the map possible. To explain how to operate the simplification, the classification and then the representation of the information, we will use different notions.

Notes

Summary

1m 27s

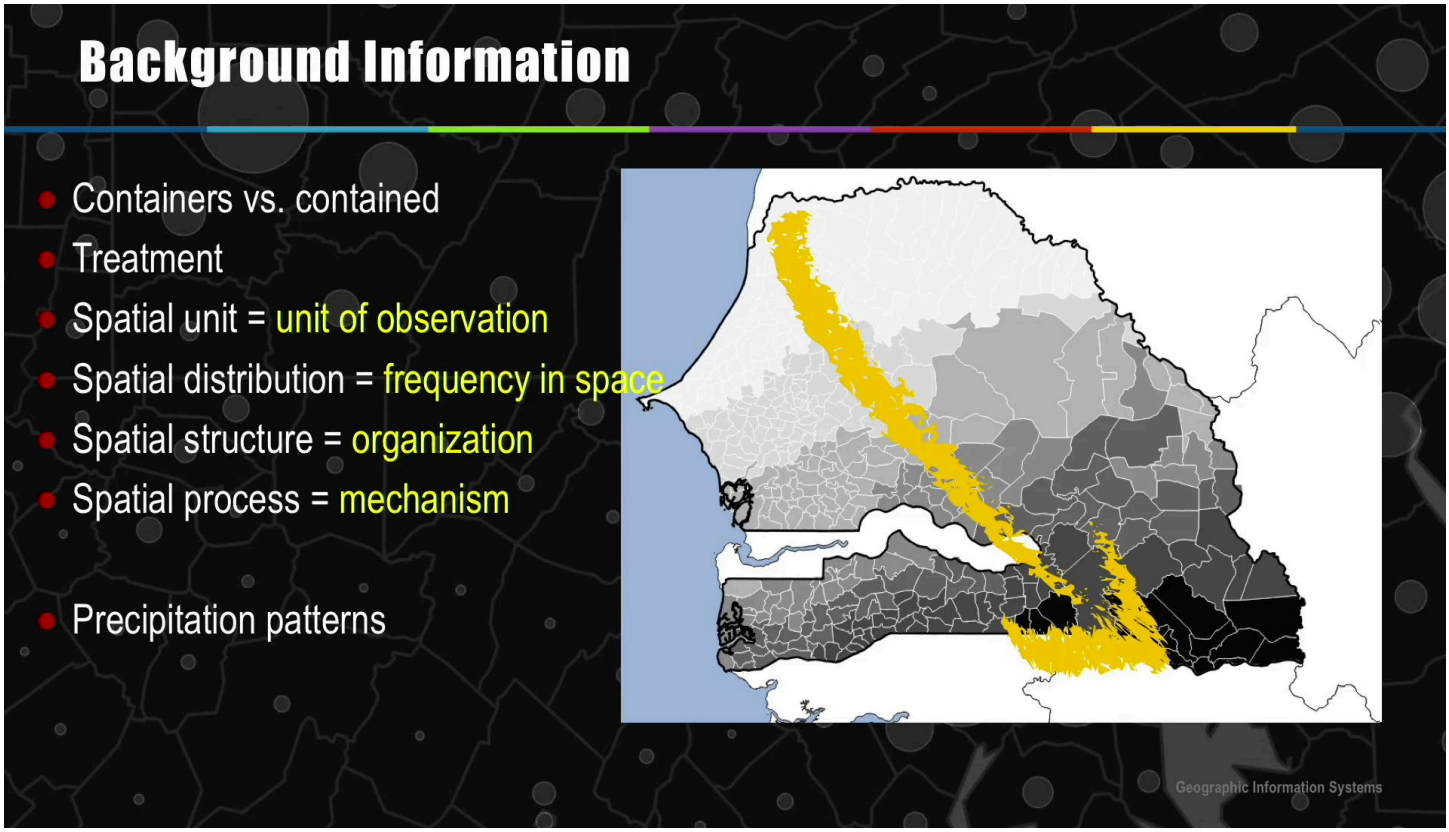


First of all, it is that of spatial unity which constitutes an indivisible and georeferenced unit of observation, such as the polygon, the point or the polyline.

Notes

Summary

2m 10s

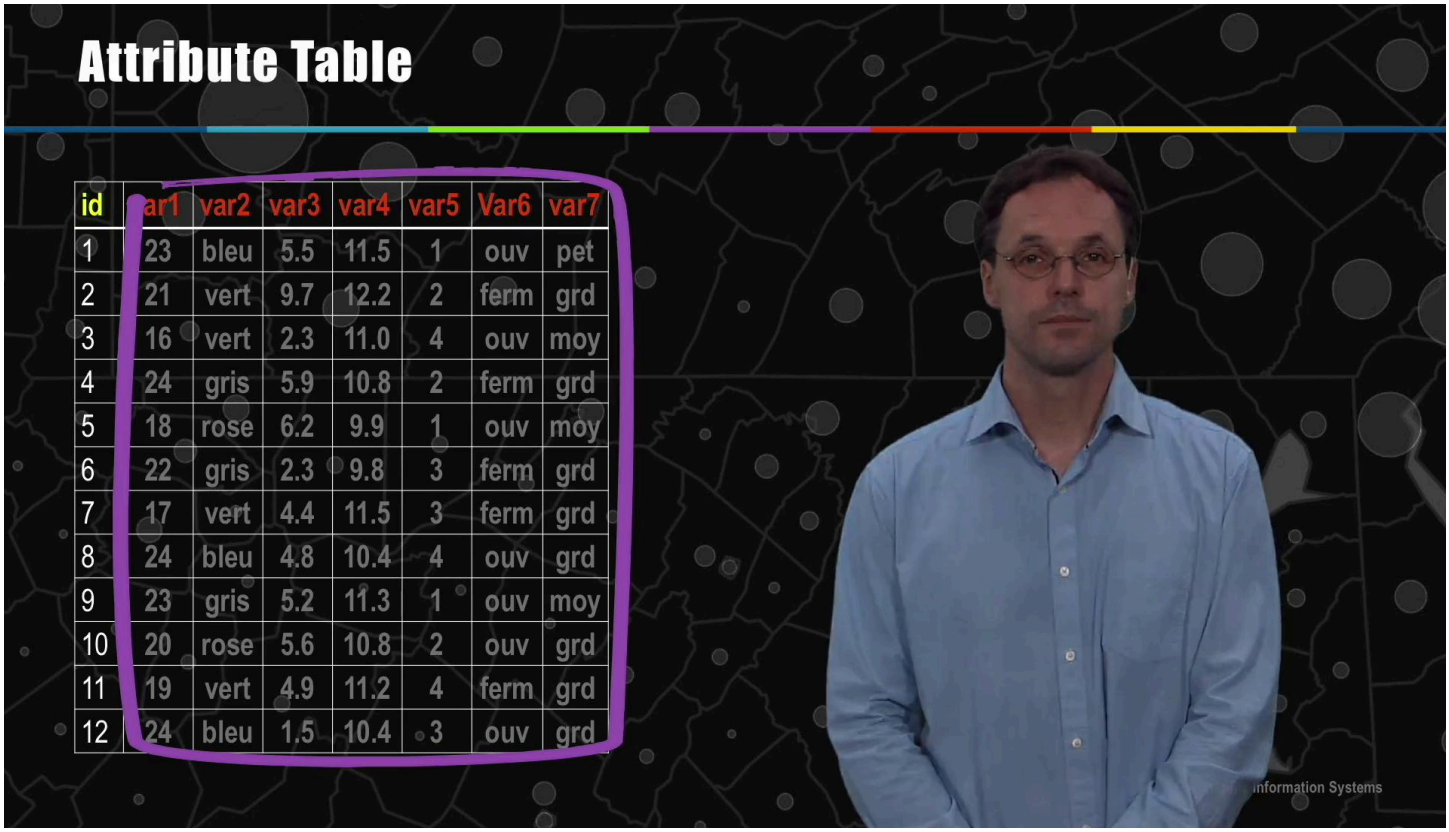


Then, the spatial distribution concerns the distribution of information. This is the frequency at which a phenomenon appears in the geographical space. The spatial structure is linked to the organization of the information, to the location of each spatial unit in relation to the others, in relation to their ensemble. Finally, what we call the spatial process relates to the mechanisms which generate the spatial structures of the distributions.

Notes

Summary

2m 21s

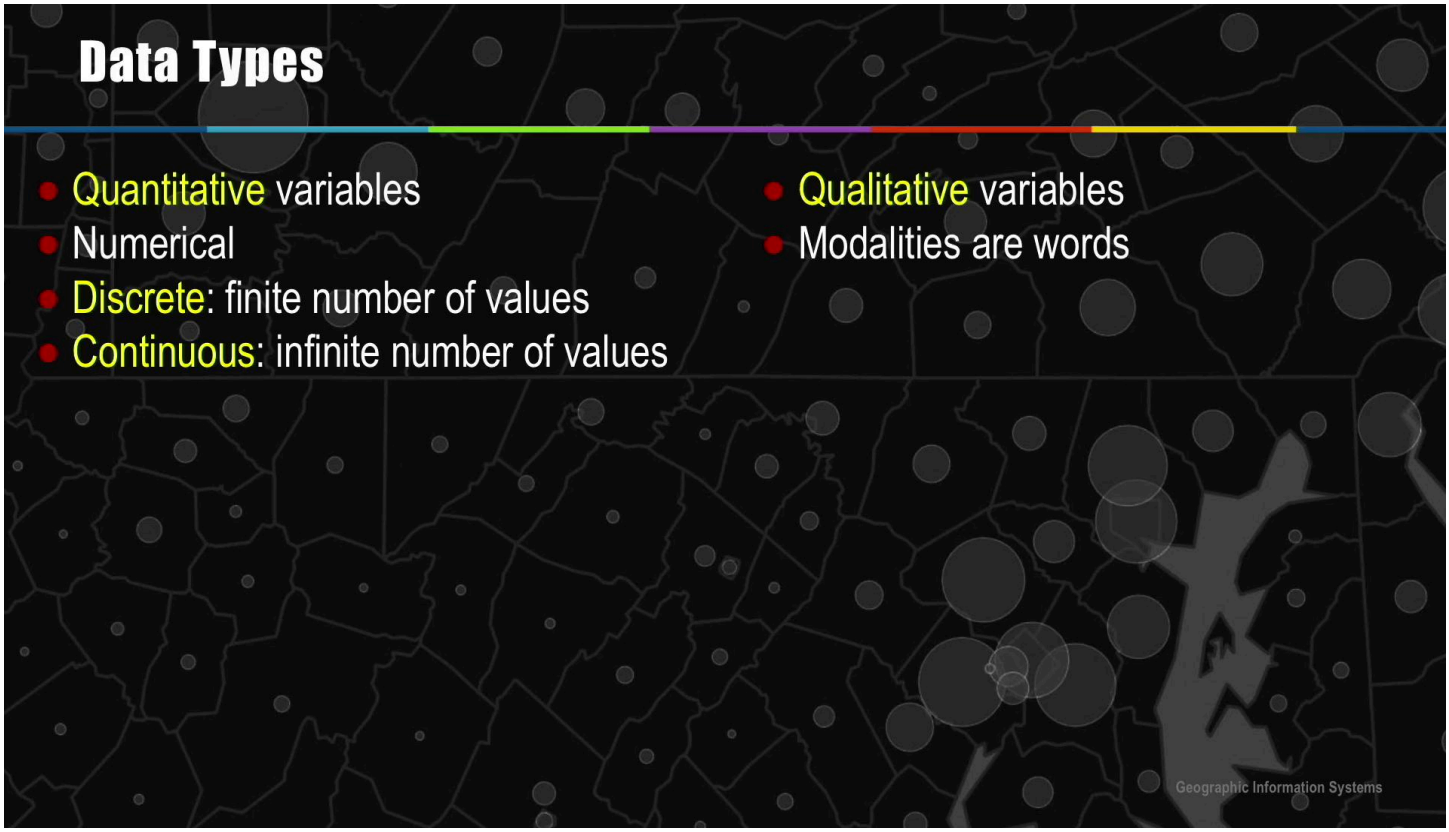


To store the information, we will use a double entry table. The horizontal lines of this table represent the spatial units studied. The vertical columns of the table are used to store the thematic attributes. The spatial units are always identified in a unambiguous way. It is a special variable which fulfills this role, a numerical identifier, here in yellow, the ID column. This digital identifier enables to ensure that each space unit is unique and it allows to make the link with its geometry stored in another file. Statistical thematic mapping will allow us to evaluate the evolution of thematic variables in space. They can also evolve in time and in this case, we will need to use several columns representing several time states. These thematic variables can be a direct measure, either a precise numerical descriptor or qualitative expert, or a derivative and indirect indicator.

Notes

Summary

2m 48s

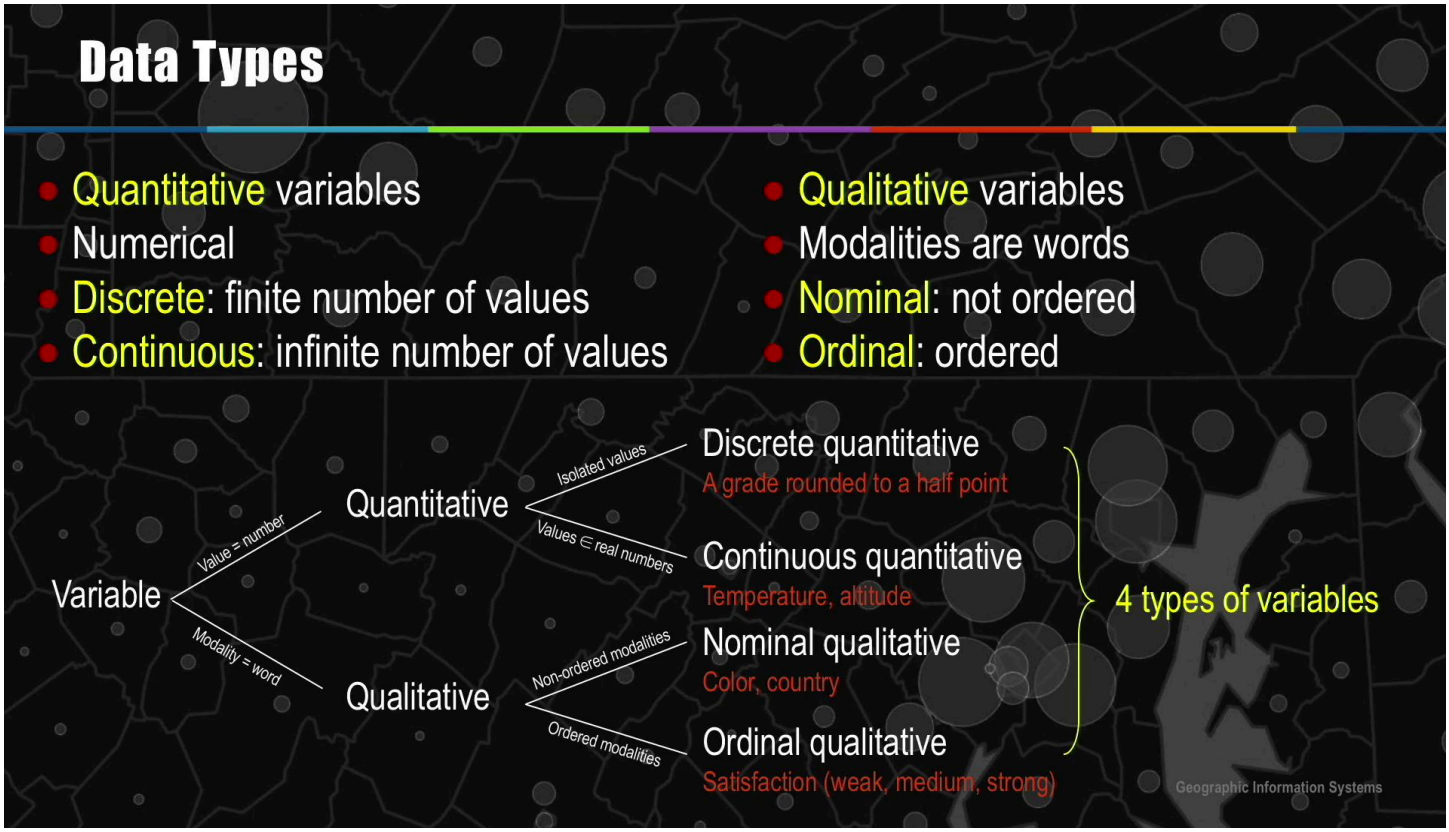


There are two main types of thematic variables, the qualitative data and the quantitative data. A variable is quantitative if the values it can take are numbers and reflect a notion of magnitude. These numerical magnitudes come from counts, measurements or calculations made on counts or measurements. The quantitative variables can be discreet or continuous. They are discreet when the information they represent takes a finite number of values between two ordinary values. The values of the discreet variables are precise, isolated and often integer numbers. The quantitative variables are continuous when the information they represent takes an infinite number of valuesbetween two ordinary values. In that case, the values are numbers from an interval of real numbers. In general, the discreet quantitative variables count and the continuous quantitative variables approximate. A qualitative or categorial variable is a variable for which the information which characterizes each spatial unit does not represent a quantity. The different valuesthat this variable can take are called the modalities.

Notes

Summary

3m 50s



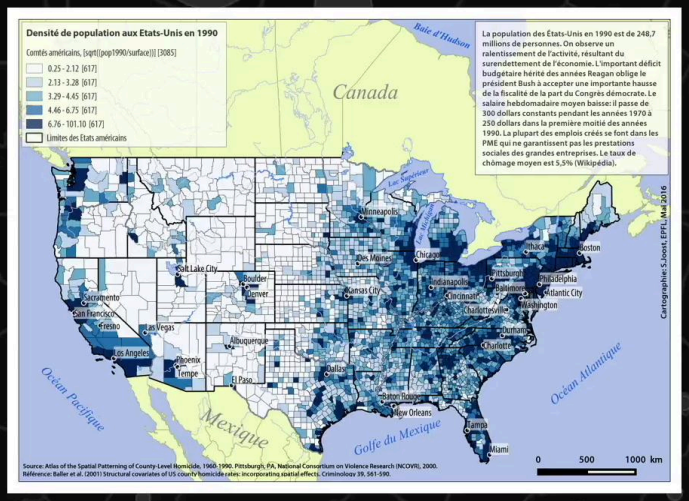
A distinction is made between the nominal qualitative variables whose modalities cannot be ordered and the ordinal qualitative variables whose modalities can be ordered according to their sense. It is necessary to always use two terms to precisely characterize a variable among the four existing types. This type can be discreet quantitative, continuous quantitative, nominal qualitative or ordinal qualitative.

Notes

Summary

5m 02s

Absolute and Relative Quantitative Data



- Absolute quantitative variable, raw value
- Visual variable = size
- Example: population size
- Proportional symbols (circles)

- Relative thematic variable
- Valid for all surfaces represented
- Visual variables = value and/or color
- Example: population density

Geographic Information Systems

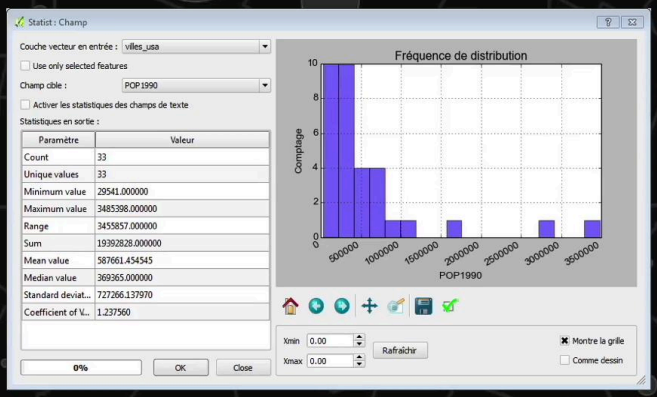
An important rule should be applied to quantitative variables: when the thematic variable is gross and that it comes from a counting or enumeration, it is mandatory to use the visual variable of size which directly represents these quantitative variations by the variations of the surface of a symbol. Thus, a number of inhabitants, a quantity of tonnes or a number of vehicles will always be represented by proportional circles for example, centered on the corresponding locality or on the centroid of a polygon. This is the subject we will discuss first. However, when a statistical variable is relative and has been transformed by a calculation as is the case on the map displayed on the screen, it is valid in all points of the area of the considered geographical units. We will therefore be able to use the graphic variable of the value and the color which is adapted to the representation of the ordered information. The averages, the rates, the indices are relative values which are always defined in relation to another value. This is the case for the population density or for the share of unemployed people in the active population. The cartographic representation of this type of variables is part of the color range maps category which we will discuss in a second step.

Notes

Summary

5m 30s

Maps with Proportional Symbols



- Mathematic relation between values and symbol area
- Know the statistical distribution
- Maximum, minimum, range
- In QGIS, "Basic Statistics" or "Statist" extenstion

Geographic Information Systems

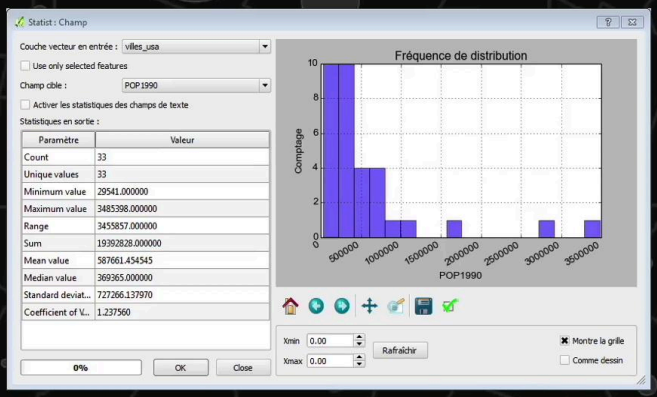
The cartographic representation of an absolute value in proportional symbols amounts to applying a mathematical relation between the statistical distribution of recorded values and the area of the proportional symbols to be represented on the map background. This is the surface of the symbols which is visualized and which constitutes the parameter which varies proportionally. Firstly, it is necessary to know the statistical distribution of the chosen variable and in particular the maximum and minimum values which define the extent. In QGIS, the vector analysis tool "basic statistics" or the "Statist" plugin allow to quickly get acquainted with the descriptive statistics of the distribution and even to visualize the dispersion of the variable in Statist. We will be able to use these values in the process of construction of proportional circles as we will see later. The extent is important because it is on this basis that we will be able to choose a ratio between the value of the variable and the area of the proportional symbol which will allow to represent correctly all the spatial units. The characteristics of dispersion and distribution are also important.

Notes

Summary

6m 53s

Maps with Proportional Symbols



- Mathematic relation between values and symbol area
- Know the statistical distribution
- Maximum, minimum, range
- In QGIS, "Basic Statistics" or "Statist" extenstion
- Weak dispersion, symbols with same size ☹
- Strong dispersion, very small and very large symbols ☹
- Extreme cases: transform the ratio

Geographic Information Systems

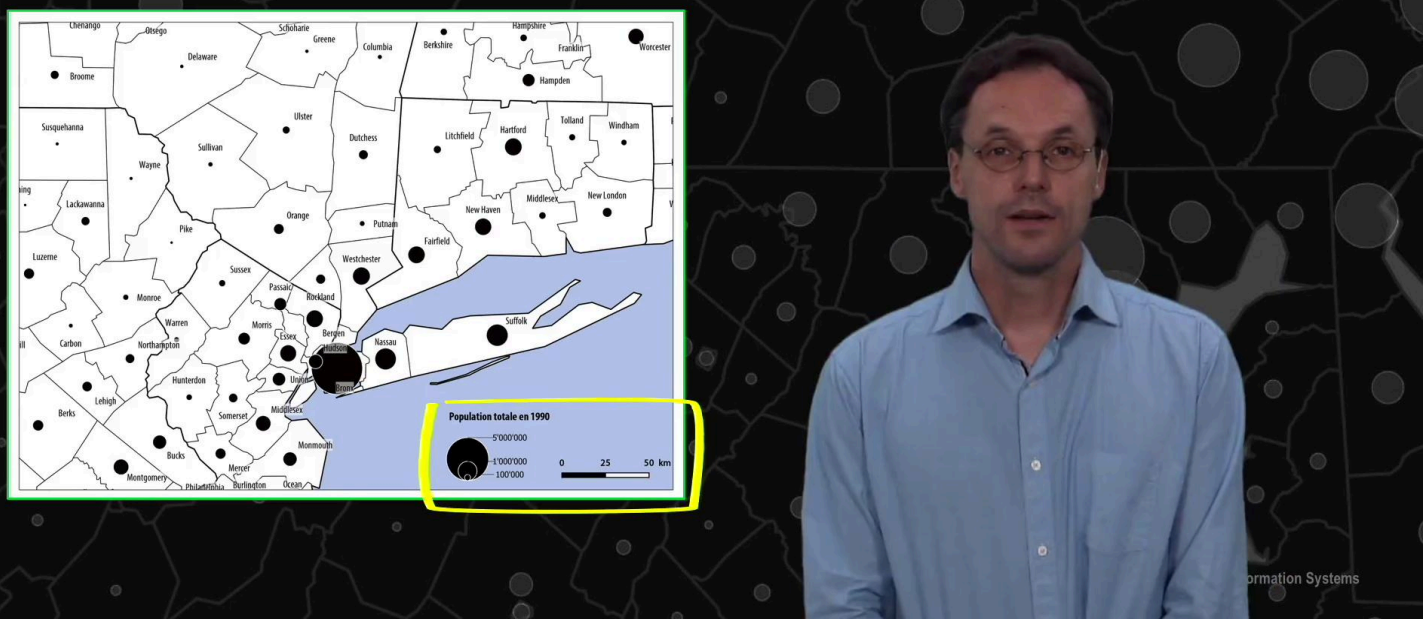
If the dispersion of the valuesis small, the risk is to build a map on which the symbols have nearly all the same size except the extremes. And in the case of a very important dispersion, the risk is to produce only small and large symbols on the map. We must be attentive to these extreme cases and if necessary, we will have to transform the value/area ratio.

Notes

Summary

8m 02s

Proportional Symbols – Spatial Constraint



The spatial distribution and the size of the spatial units are also a constraint because these elements influence the distribution of points which will support the symbols in the plan. The major problem to solve is to size the largest symbols so they do not hide the map background and so that their hold does not crush the smallest symbols. Then, a large density of spatial units like here with the total population in the New York area will compromise the readability of the map especially when, due to its proximity to an urban center, the values it shows are all high. On the other hand, the big cities in the center of agglomerations have values far superior to those of spatial units which surround them. And to represent on the same map these very different and very close values is a real problem. In our New York example, the level of transparency can help and allow a better readability. But in this kind of case, the only satisfactory solution is the decomposition of the map into two views, one showing a general representation of the region and another that zooms on the zone for which the first view is almost unreadable and does not provide any information. It is then necessary to adapt the rule of proportionality to represent the information in the area where the zoom is done.

Notes

Summary

8m 25s

Proportional Symbols – Value-Area Ratio

- Weakly distributed variable

id	var1	var2	var3	var4	var5	Var6	var7
1	23	bleu	5.5	11.5	1	ouv	pet
2	21	vert	9.7	12.2	2	ferm	grd
3	16	vert	2.3	11.0	4	ouv	moy
4	24	gris	5.9	10.8	2	ferm	grd
5	18	rose	6.2	9.9	1	ouv	moy
6	22	gris	2.3	9.8	3	ferm	grd
7	17	vert	4.4	11.5	3	ferm	grd
8	24	bleu	4.8	10.4	4	ouv	grd
9	23	gris	5.2	11.3	1	ouv	moy
10	20	rose	5.6	10.8	2	ouv	grd
11	19	vert	4.9	11.2	4	ferm	grd
12	24	bleu	1.5	10.4	3	ouv	grd

- Identify the maximum value = 24
- An 8mm radius = ~200 pixels^2
- $A = \pi \cdot r^2$ and $r = \sqrt{\frac{A}{\pi}}$
- 24→A =200 and r=8;
- 23→ A =190 and r=7.7;
- 21→A =174 and r=7.4;
- 20→A =166 and r=7.2, etc.

Geographic Information Systems

If the variable to be represented is not too dispersed as is the case here with variable 1, we will identify the maximum value to represent to determine the size of the largest symbol and determine on this basis the ratio that will enable to represent all the valuesin proportional circles. If we want to represent a radius of 8 millimeters for the maximum value, we know that the corresponding surface in square pixels is about 200. The surface is indeed equal to π multiplied by the square radius. We therefore use the ratio 24 to 200, then a rule of three to calculate the surface and the radius of the other symbols in simple proportionality.

Notes

Summary

9m 48s

Proportional Symbols – Value-Area ratio

- Strongly dispersed variable

id	var1	var2	var3	var4	var5	Var6	var7
1	23	bleu	5.5	11.5	1	ouv	pet
2	21	vert	9.7	12.2	2	ferm	grd
3	16	vert	2.3	11.0	4	ouv	moy
4	24	gris	5.9	10.8	2	ferm	grd
5	18	rose	6.2	9.9	1	ouv	moy
6	22	gris	2.3	9.8	3	ferm	grd
7	17	vert	4.4	11.5	3	ferm	grd
8	24	bleu	4.8	10.4	4	ouv	grd
9	23	gris	5.2	11.3	1	ouv	moy
10	20	rose	5.6	10.8	2	ouv	grd
11	19	vert	4.9	11.2	4	ferm	grd
12	24	bleu	1.5	10.4	3	ouv	grd

- Ratio based on the most unadapted value
- Transform the ratio
- Logarithm or square root
- Loss of the simple link between symbol's values and areas
- Mention the type of transformation in the legend

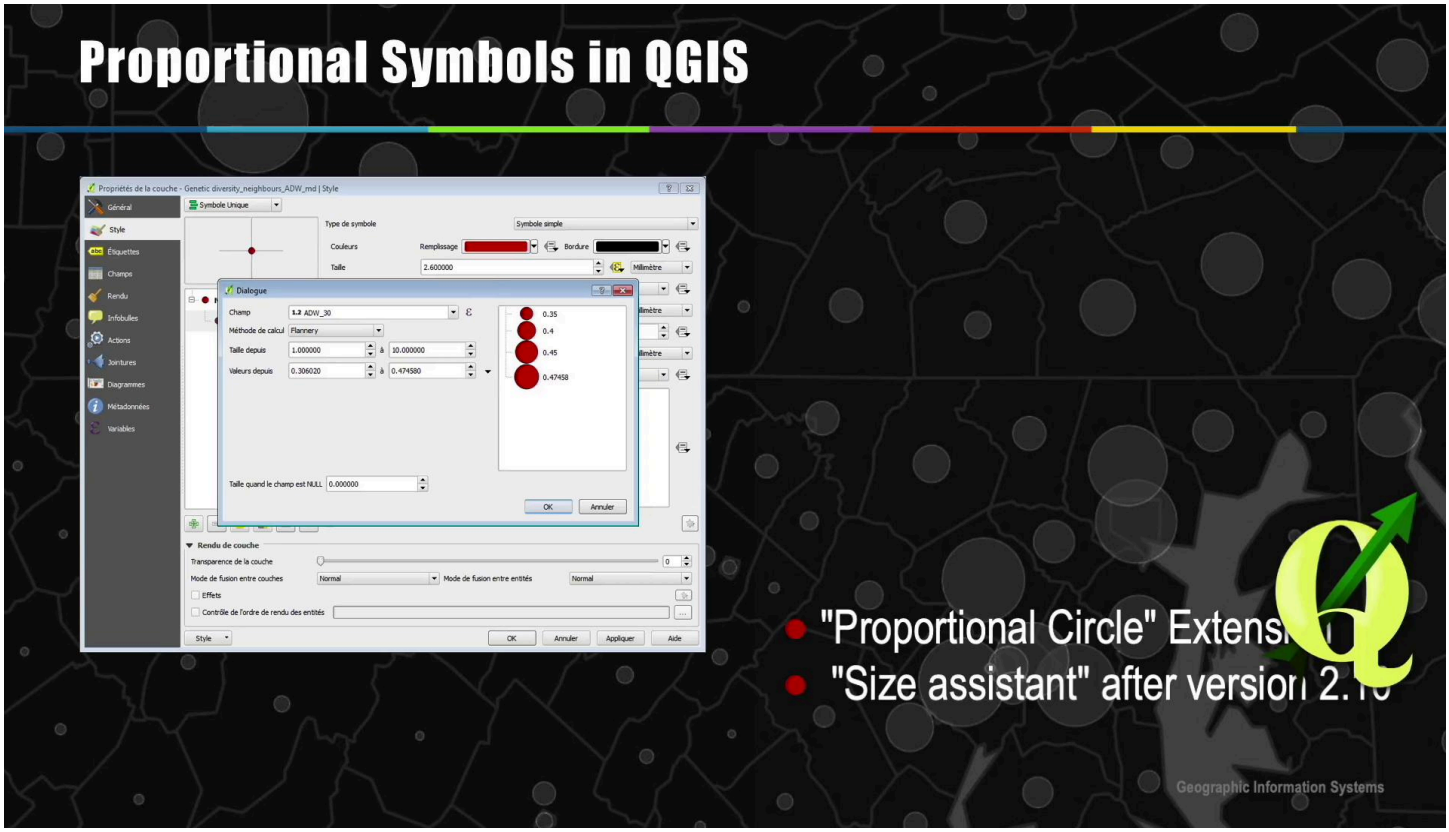
Geographic Information Systems

Now, if the variable to be represented is too dispersed, the surface of the smallest symbol resulting from the choice of ratio determined from the highest value risks to be inadequate. In that case, it is necessary to transform the ratio to make it go from a simple and linear ratio such as "value x ratio = surface", to a more complex ratio, logarithmic or root, such as "log value x ratio = surface" or "root value x ratio = surface". We then lose the simple connection between the value and the surfaces of the symbol with the risk that the map reader finds it difficult to understand the transformation when he expects a simple proportionality. That is why we must always indicate the type of transformation in the key when we use it.

Notes

Summary

10m 28s

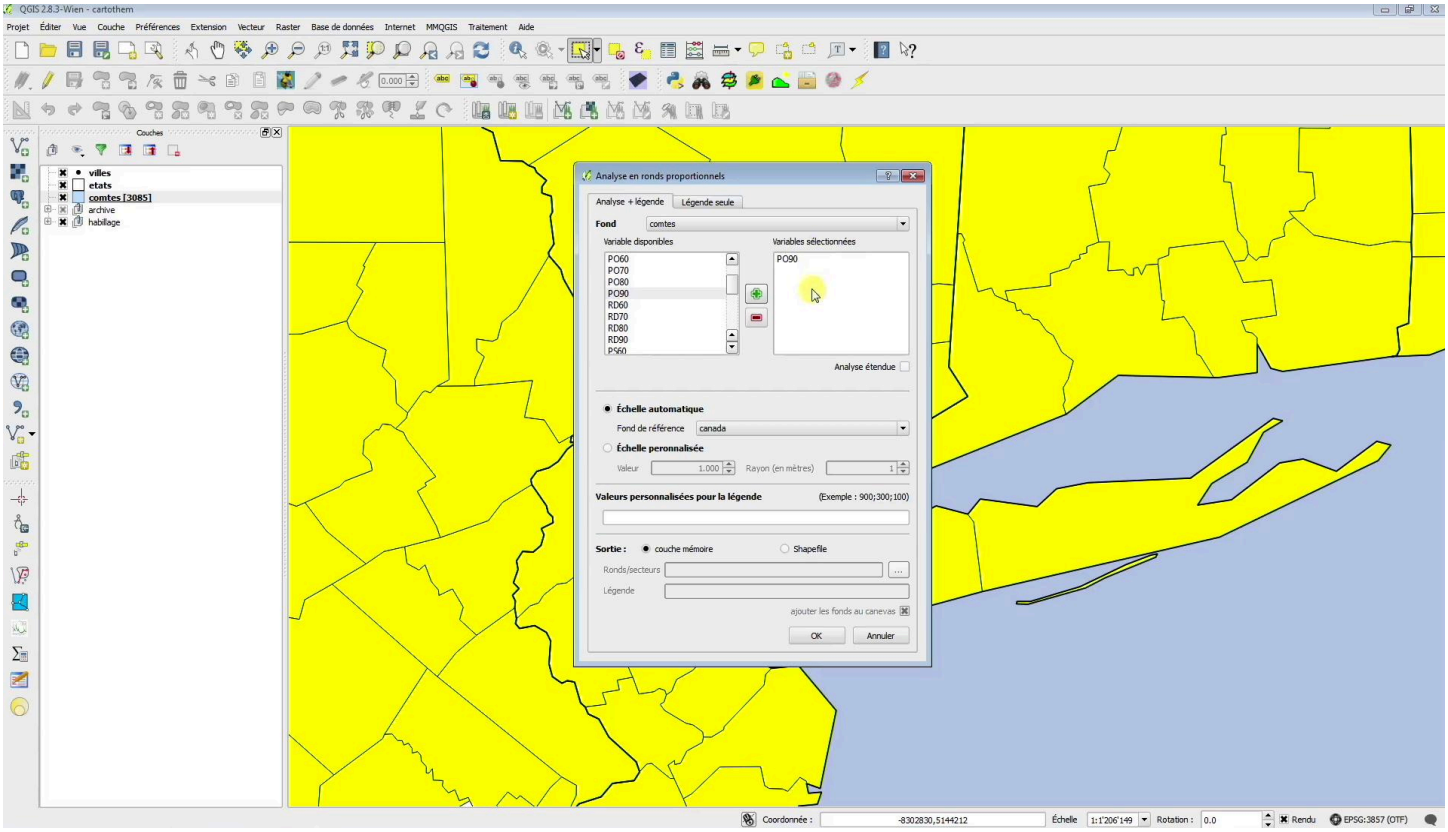


In the version 2.8 of QGIS that we use in this MOOC and which is a long-term maintained version, it is the "analysis in circles and proportional sectors" extension which enables to create maps in proportional circles. We will use the latter in the following example, but for your information, know that from version 2.10 of QGIS, a function called "size assistant" is integrated and accessible via "the properties of the layer" in the "style" tab, under "single symbol", then clicking on the "size parameter expression" icon. So let's now look in QGIS how the analysis in proportional circles works.

Notes

Summary

11m 15s

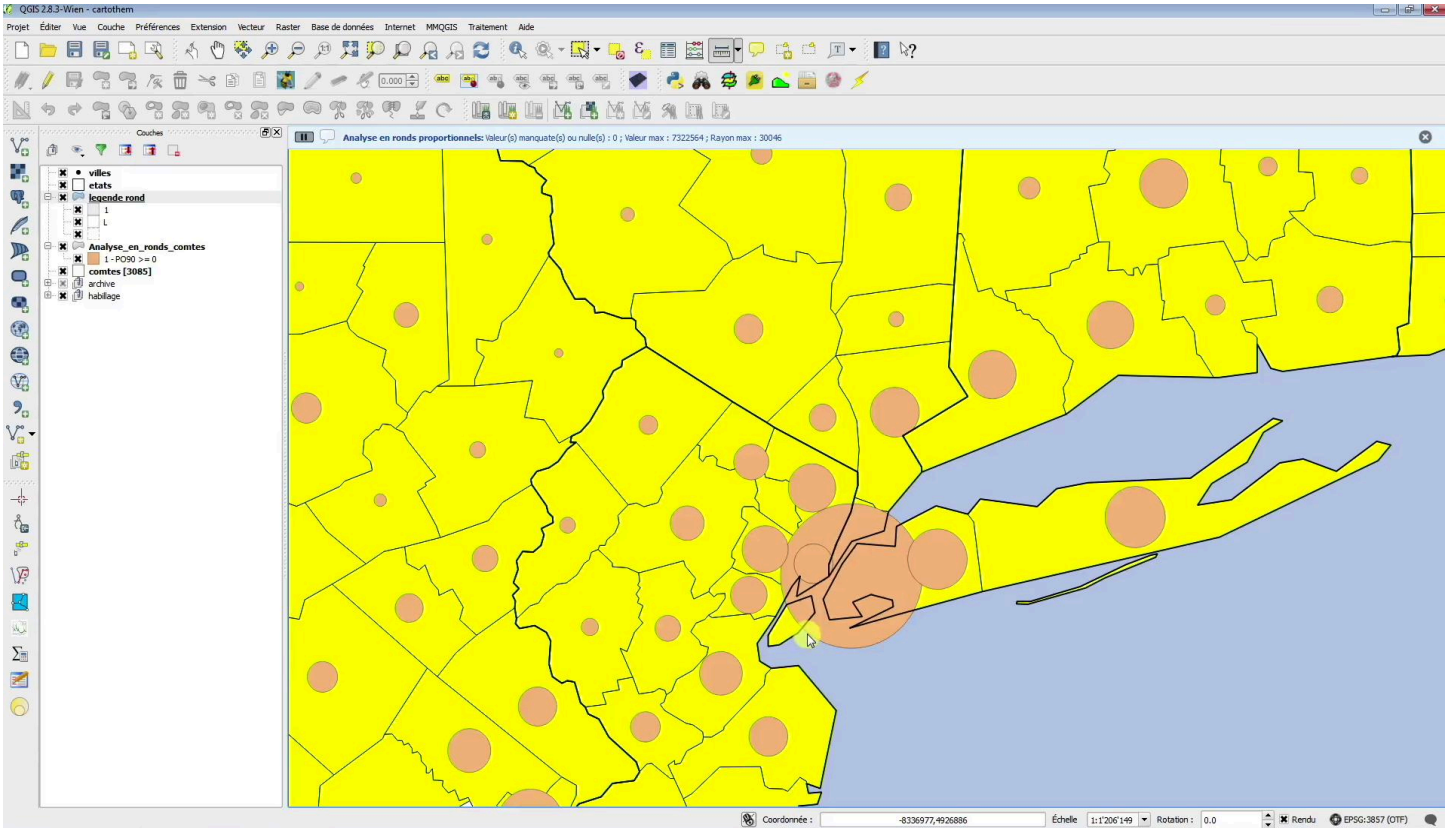


Via the "extensions" menu, "install/manage the extensions", we have previously installed the "analysis in circles and proportional sectors" plugin. On the map of the american counties, we will zoom in on the New York area to represent the total population by county, expressed in number of inhabitants. As we saw earlier, it is important to know the distribution of the values of variables to be represented, in particular the maximum value and the minimum value. It is possible to obtain these information by selecting the counties in question on the screen, then by going to the "vector" menu, under "analysis tools", then "basic statistics". We select the layer of counties by specifying that we would like to use only the valuesof the desired field for the selected geographical units, that is to say the total population in 1990, then we click on OK. The maximum value is 7,300,000 and the minimum value is about 28,000. We can now launch the extension by clicking on the icon representing proportional circles or via the "vector" menu, then "analysis in proportional circles". In the "analysis + key" tab which interests us, it is necessary first of all to select the layer of counties, then the population variable in 1990 and add it to the list of selected variables.

Notes

Summary

11m 55s

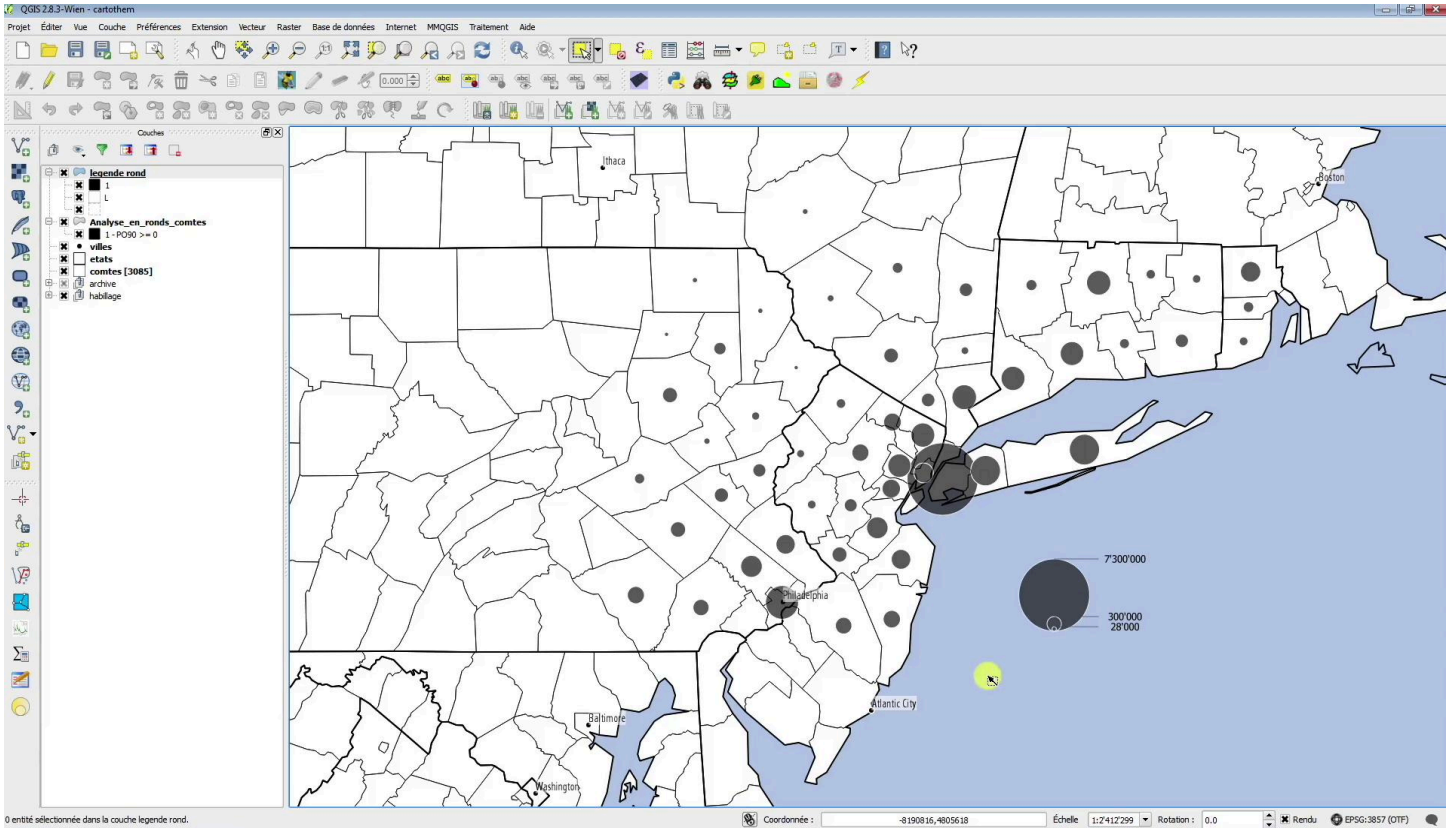


Then, we click on "custom scale", an option that asks to provide on one hand the maximum value to be represented and which is 7,300,000, and on the other hand the radius corresponding to this maximum value and which must be represented on the map. This value must be given in meters. We can use the "measure the length" tool which will help to determine this radius. In our case, 30,000 meters seems to be a suitable value and we can transfer it to the corresponding field. Then, in the "custom values" field for the key, it is possible to define a list of values which we want to see appear in the key. Values must be separated by a semicolon or a space. If the box is left empty, three values will be calculated by default: the maximum, the maximum divided by three and the maximum divided by nine. We will ask to display the maximum, so 7,300,000 the median which is of 300,000 and the minimum. Finally, we can choose whether the layer of proportional circles and the corresponding legend must be stored only in the memory of QGIS or if we want to create two files in shape format to use them again later in another session. In this last case, it will be necessary to indicate the directory in which to save the shape files.

Notes

Summary

13m 28s

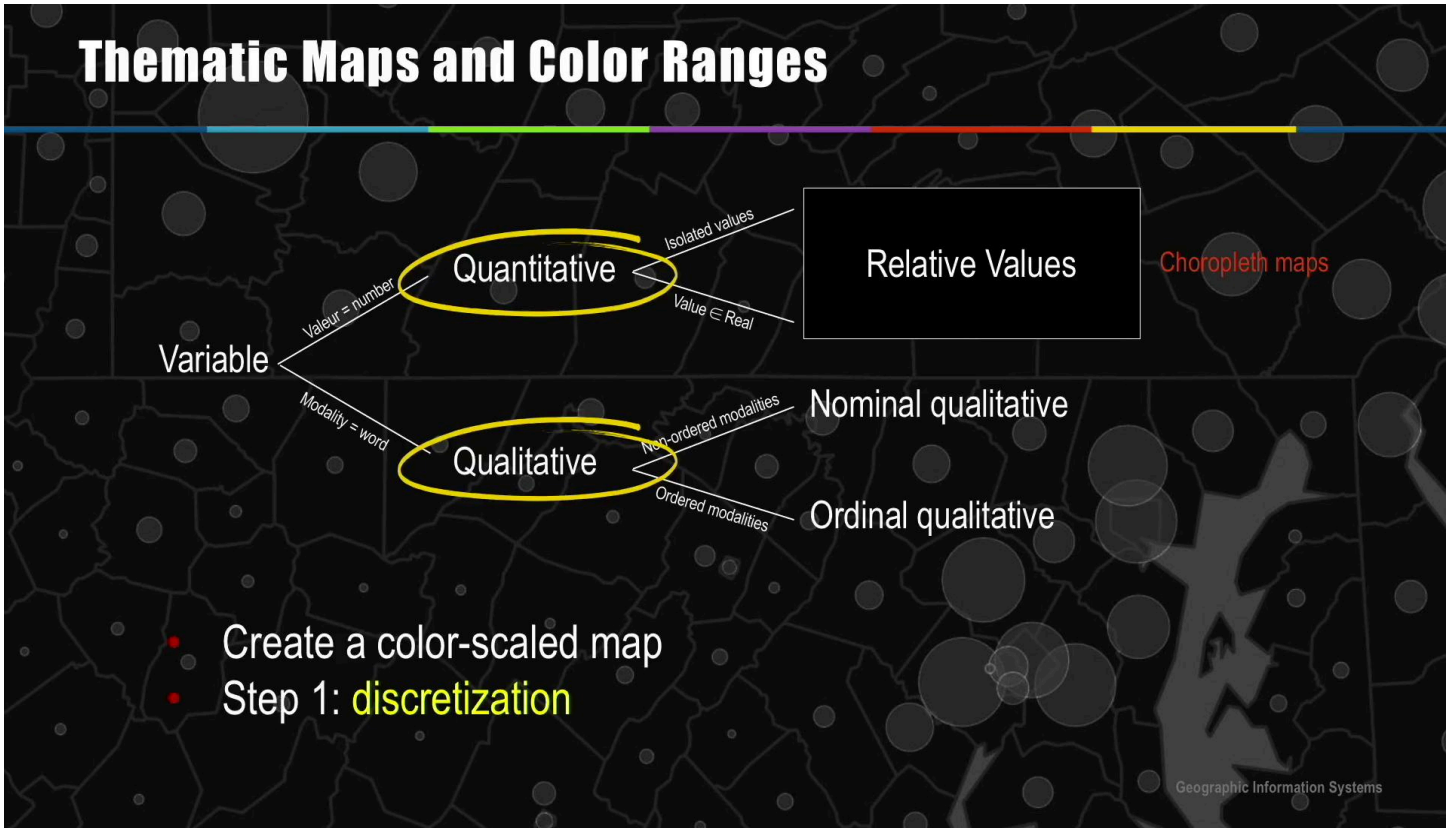


Here we choose the "memory layer" option which requires the "memory layer saver" extension so that the layers of information can be recorded. We then click on OK to generate the proportional circles and the key. We will now move the key and the proportional circles above the other layers, then adjust their colors. A suggestion is to use black for filling, the white for the outline, then to use the transparency at 40% in order to identify the limits of the underlying layers, and an edge width of 0.3. The same parameters are then applied to the key, filling in black, outline in white, border of 0.3 and transparency of 40%. Finally, it is possible to move the key. It is necessary to select the corresponding layer, here "circles key", then select the key on the map, then make it editable and use the "move entity" tool to place it in the desired location. Finally, save the changes. Here ! You now know how to produce a map in proportional circles. You will learn in the following lesson how to retrieve key elements for the layout of a printing dialer.

Notes

Summary

15m 00s

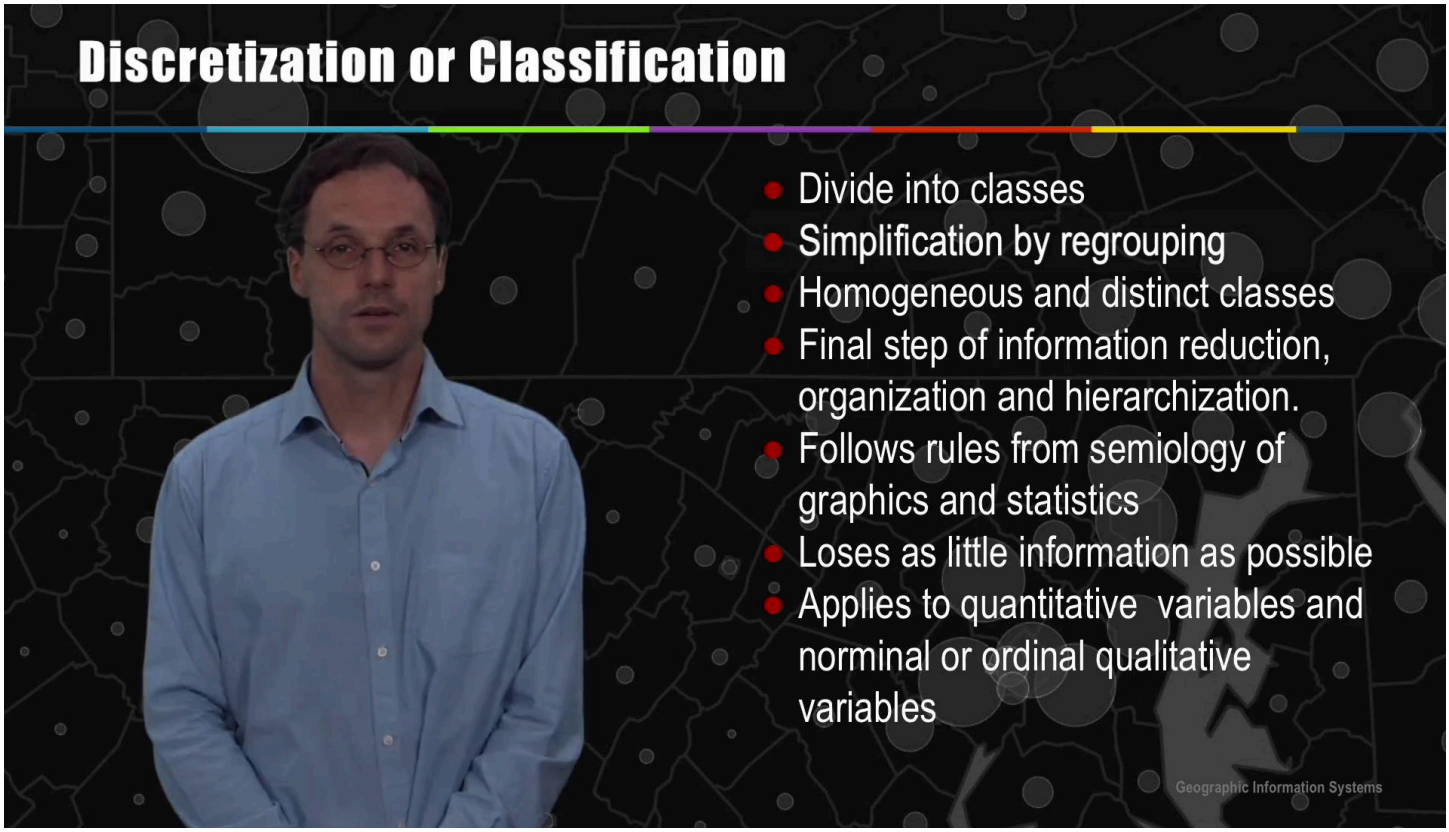


We will now move on to the elaboration of thematic maps in color ranges. This category is divided into two groups, on one hand the group that concerns the representation of qualitative nominal and ordinal data and on the other hand, the group dealing with relative quantitative data. In the latter case, we speak of choropleth maps, a term that will be explained later. To create a map in ranges of colors with a data set composed of spatial units and their thematic attributes, it is necessary to proceed, firstly, to what is called the discretization.

Notes

Summary

16m 35s



The discretization is the operation that allows to cut into classes the statistical distribution of the values of a qualitative variable or quantitative variable, in order to simplify the information by grouping geographical objects presenting the same characteristics in distinct classes. The discretization must allow to create homogeneous classes and distinct from one another. The geographical objects of the same class must resemble each other more than they resemble objects of other classes. The discretization constitutes the last step of the reduction, the organization and the prioritization of the information and it allows to create a map which accounts for the spatial distribution of a statistical distribution. The class setting operation must satisfy the requirements of the cartographic representation and that of the statistical principles. It must retain the essential characteristics presented by the data and lose as little information as possible, but also respect the rules of the graphic semiology that enable to optimize visual perception and transmit effectively a geographic information of quality. It is applicable to quantitative and qualitative variables.

Notes

Summary

17m 09s

Discretization of Qualitative Variables

- Nominal: generalizes information according to implicit hierarchy
- Maximizes discrimination between spatial units

Geographic Information Systems

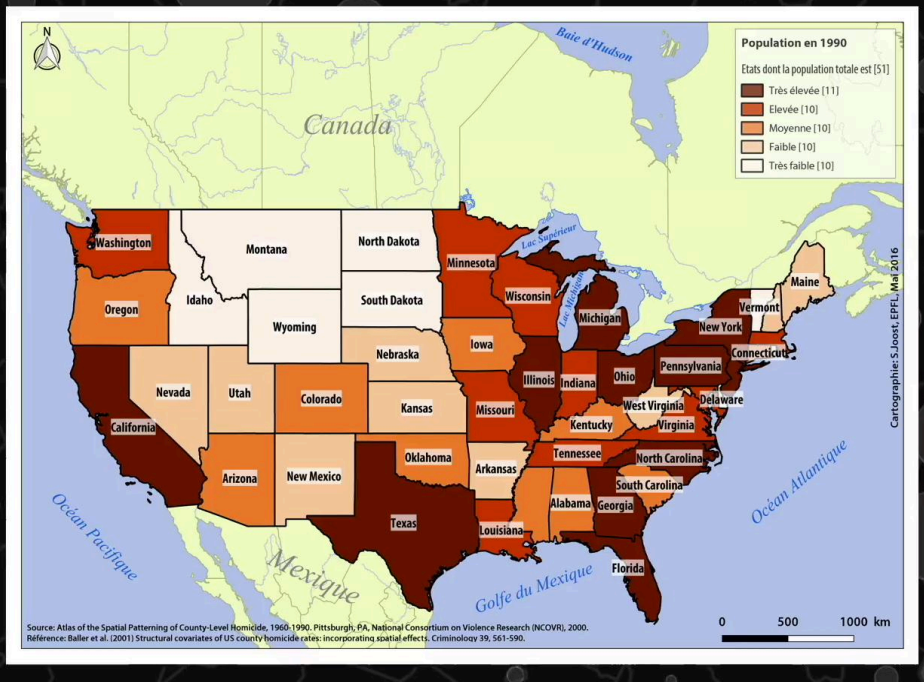
In the case of nominal qualitative variables, all the modalities can be attributed to different classes in order to generalize the information if a hierarchical structure exists. In the example on the screen, we have classified the american counties according to their belonging to the states, but the counties could have also been divided into four classes: the North-West, North-East, South-West and South-East states. Nominal qualitative charts do not require special treatment except the allocation of discreet colors to differentiate at best the spatial units displayed. The main quality of this map is to allow the distinction between the states and it would be wrong in this case to resort to a representation using the variation of the value of a single shade. Here, it is a function of random assignment of colors proposed by QGIS which was used given the large number of units. Ideally, this representation should have been reused and the colors of the counties should have been changed in regions where differences are tenuous, like between Indiana, Ohio and Kentucky or between Oregon and Idaho.

Notes

Summary

18m 20s

Discretization of Qualitative Variables



- Nominal: generalizes information according to implicit hierarchy
- Maximizes discrimination between spatial units
- Ordinal: Based on ranking

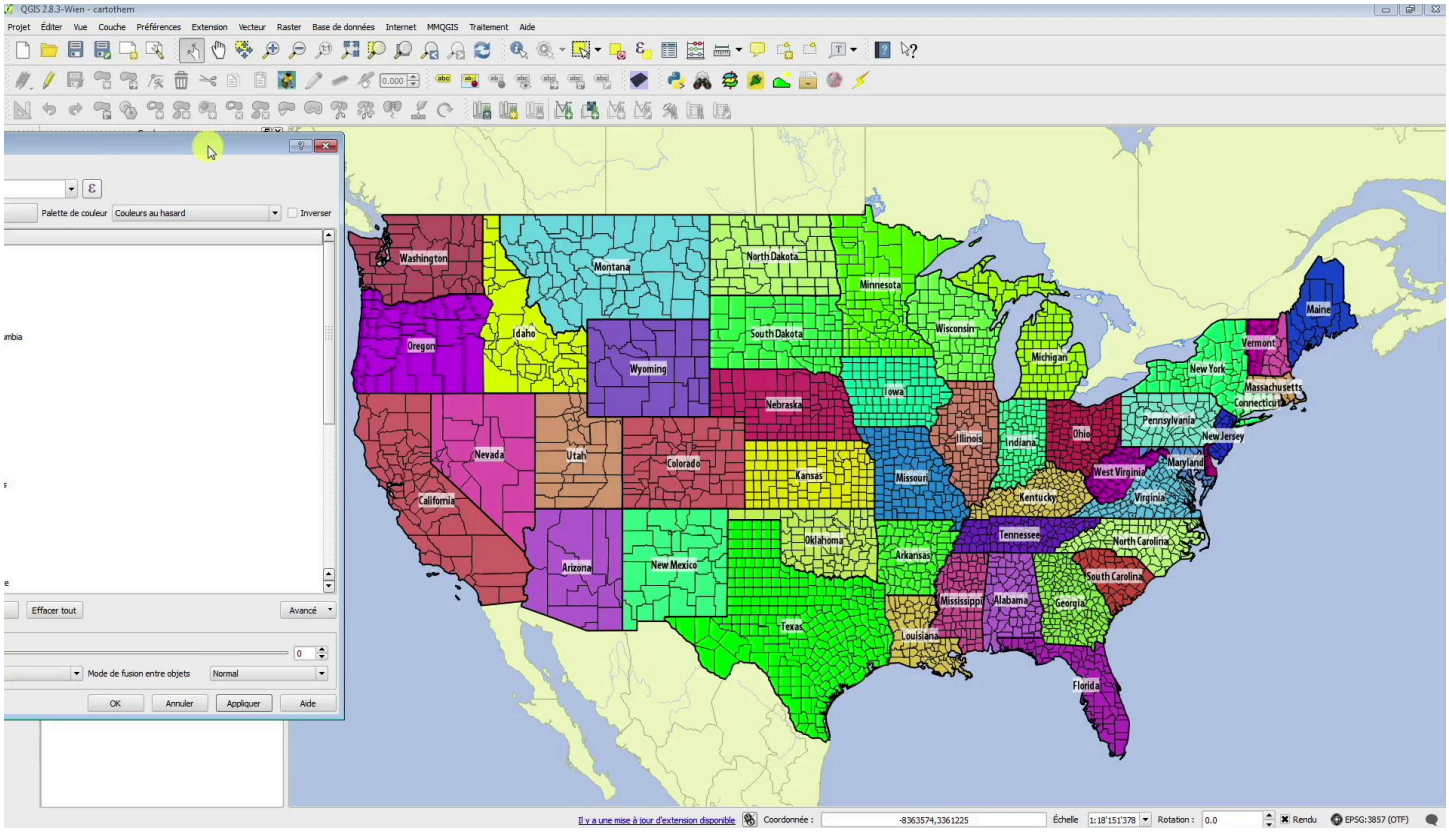
Geographic Information Systems

In the case of ordinal qualitative variables, as here the indications on the size of the total population, the difference between two successive modalities is not quantifiable and it always corresponds to a unit. But it is possible to assign a rank to the different values and to discretize the distribution in question on this basis. We will now go to QGIS to explain to you how to produce a nominal qualitative map.

Notes

Summary

19m 30s

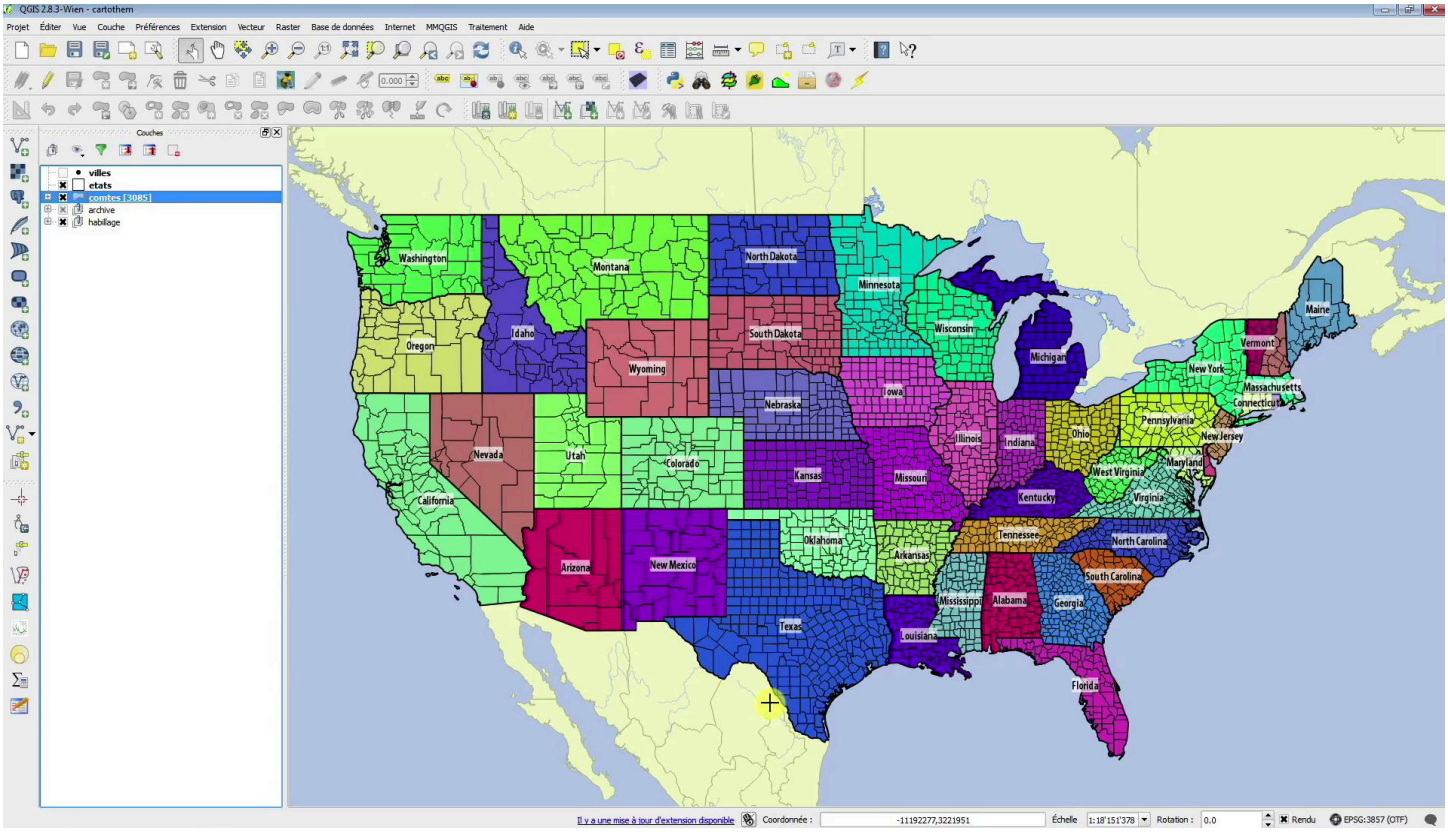


In QGIS, a project has already been opened containing in particular the layer of american counties which we select here. A double-click on the layer enables to reach the properties of the layer and the "style" tab. In the drop-down menu, located at the top of the window, choose "categorize", then select the "state name" column and finally click on the "sort" button without modifying the current color palette. QGIS will then list all the states contained in the "county" layer and assign to each of them a shade according to the preselected palette, here gradients of blue. Our concern is to assign distinct colors to all the counties in order to be able to differentiate the states they constitute. On the other hand our concern is to be able to count on an automatic filling function since the number of units considered is significant and that it would be tedious to undertake this task manually. We will therefore choose an appropriate color palette, so "random colors". By clicking on the "apply" button, we can get an idea of the way the different colors have been distributed over the counties.

Notes

Summary

19m 57s



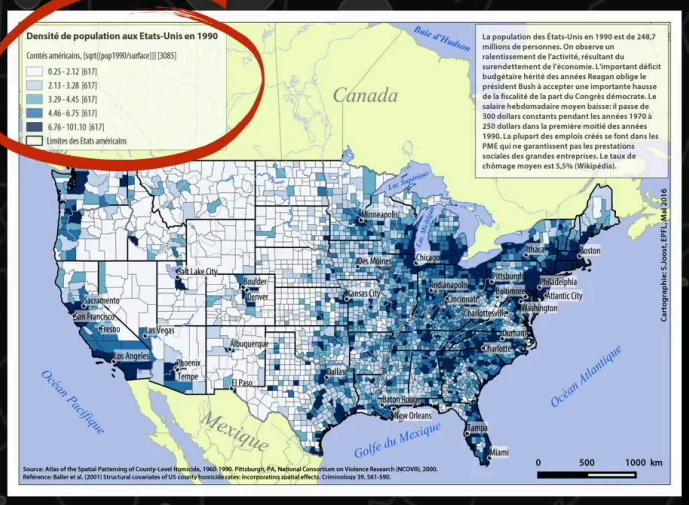
if the color distribution is not good, so if it is difficult to differentiate certain states because of colors that are too close, it is necessary to re-launch the process by choosing first any other color palette as a reset, then applying again the "random color" palette and so on until a correct configuration appears, this before validating definitively the choice of this palette. It is possible to save it afterwards for a later use by clicking on the "style" drop-down menu, then "save style" and finally, "QGIS layer style file". It is necessary to determine a directory where to save this.QML file and give it a name, for example "counties_nominal.qml". This file will be usable via the "style" drop-down menu, then "load style". Finally, to validate the choice of colors, click on OK.

Notes

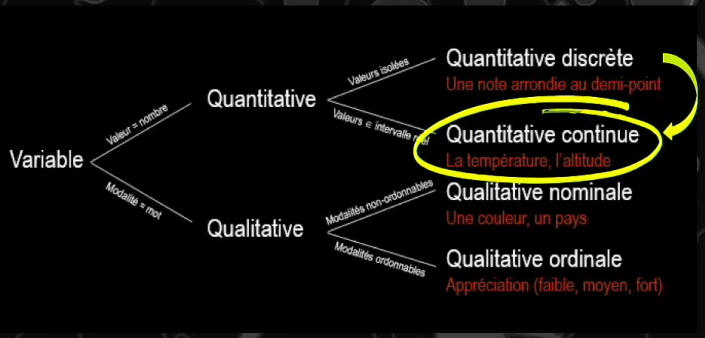
Summary

21m 14s

Choropleth Maps



- Plethos = quantity
- Khorê = ~ geographic area
- Counts, measurements
- Transformed variables
- Continuous quantitative distribution



Geographic Information Systems

Other category of cartographic representation in ranges of colors, the choropleth map is the most frequent type of statistical maps adapted to the representation of relative quantities characterizing geographical areas by means of a scale of graduated values of shades. By relative quantity, it should be understood that these are variables such as counts or measurements which have been transformed by a calculation to give ratios or percentages, distributions which then become continuous. For example, for the map displayed here on the screen, the population density per county which is a discreet quantitative variable, has been transformed. The original value has become the square root of the ratio between the number of inhabitants per county in 1990 and the surface of the corresponding county, so a continuous quantitative variable. The class setting and choice of shades and values are two crucial operations to develop such maps.

Notes

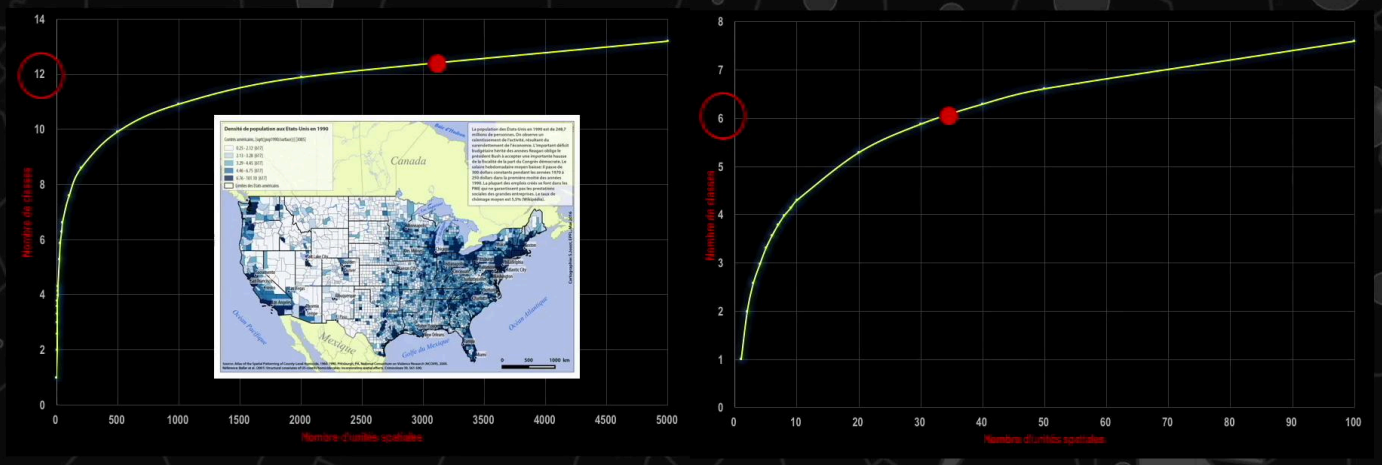
Summary

22m 21s

Discretization of Relative Quantitative Variables

$$N_{cl} = 1 + 3{,}3\, log10(N_{obs})$$

- Huntsberger index



The optimal number of classes to determine is often presented as a function of the number of spatial units in a data set. In this line of thinking, there are indices allowing to calculate the ideal number of classes for a given distribution. An example is the Huntsberger index for which the ideal number of classes, N cl is equal to 1 more 3.3 times the base 10 logarithm of the number of spatial units. If we consider a number of spatial units comparable to the example of the 3,080 american counties this formula recommends a discretization into 12 classes, which raises a problem of perception because our visual system is not able to distinguish as many perceptive thresholds, 11 in this case, in the distribution of the value of a single shade. We remember that ideally, we should not exceed six perceptual thresholds, so seven classes, in order to guarantee the maximum efficiency of the map. In that case, we see on the graph on the right that we could not theoretically create choropleth maps representing more than 33 spatial units.

Notes

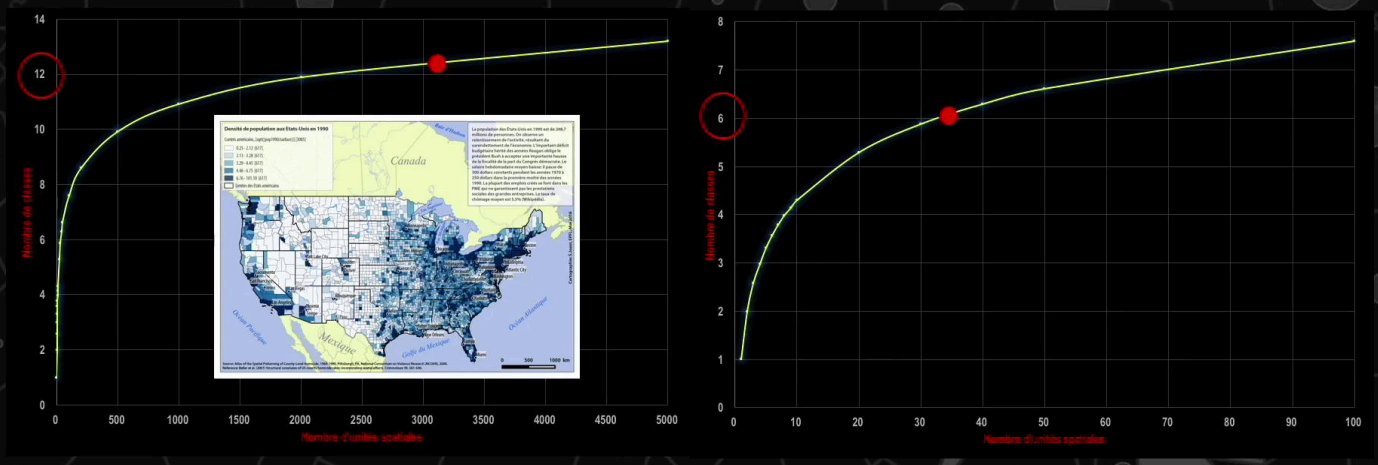
Summary

23m 22s

Discretization of Relative Quantitative Variables

$$N_{cl} = 1 + 3{,}3\ log10(N_{obs})$$

- Huntsberger index



Geographic Information Systems

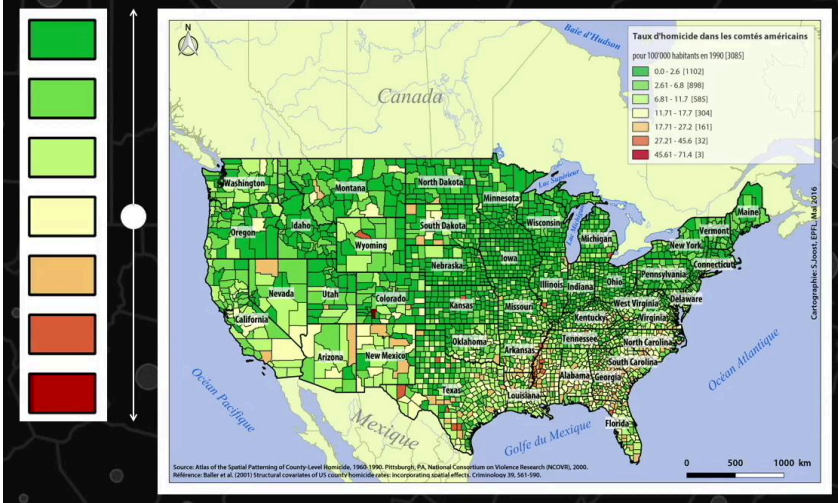
We have to consider this type of index as an indication when we work with a small number of polygons, but it is clear that the number of ideal classes is around 5 in most cases and taking it from there this is the theme represented, the characteristics of the variable and its spatial distribution which will guide the choices of the author of the map. For example, should we use an even or odd number of classes? The odd number will be chosen when we want to refer to a class whose behavior is moderate. For data sets containing a large number of spatial units, it is possible to adopt a pragmatic approach which allows to plan a greater number of classes and at the same time more perceptual thresholds.

Notes

Summary

24m 28s

Number of Classes – Large Number of Spatial Units



- Increase the number of perceptual thresholds
- Two hues
- One intermediary class
- Odd number of classes
- Theoretical maximum: 12-1 classes
- Color choice that is coherent with the theme of the map

Geographic Information Systems

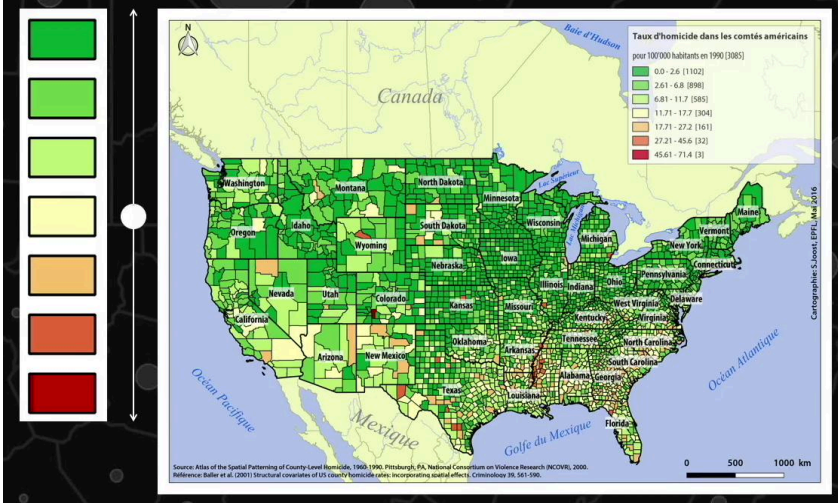
The idea is to increase the number of exploitable perceptual thresholds using a gradient of valuesfor two shades. The latter are separated by an intermediate class, represented by a neutral shade such as pale yellow for example. We will use one of the shades to represent the lowest values and the other shade for the highest values, the two being articulated around the neutral tint. We will therefore preferably use an odd number of classes, except in cases where the variable being processed represents two opposite behaviors, such as the acceptance or refusal of a voting object. Here, an even number of classes fits perfectly. We can then increase the number of classes up to a theoretical maximum of 12 based on the indications of the Huntsberger formula. We will therefore use 11 classes to obtain an odd number of classes with an intermediate class and two gradations of value in five classes, or even nine or seven classes, in order to produce the most effective map possible. This process is delicate and it is necessary to explain the double gradation system to the reader and the choice of colors must be coherent with the theme presented.

Notes

Summary

25m 09s

Number of Classes – Large Number of Spatial Units



- Increase the number of perceptual thresholds
- Two hues
- One intermediary class
- Odd number of classes
- Theoretical maximum: 12-1 classes
- Color choice that is coherent with the theme of the map
- Required explanations
- False map

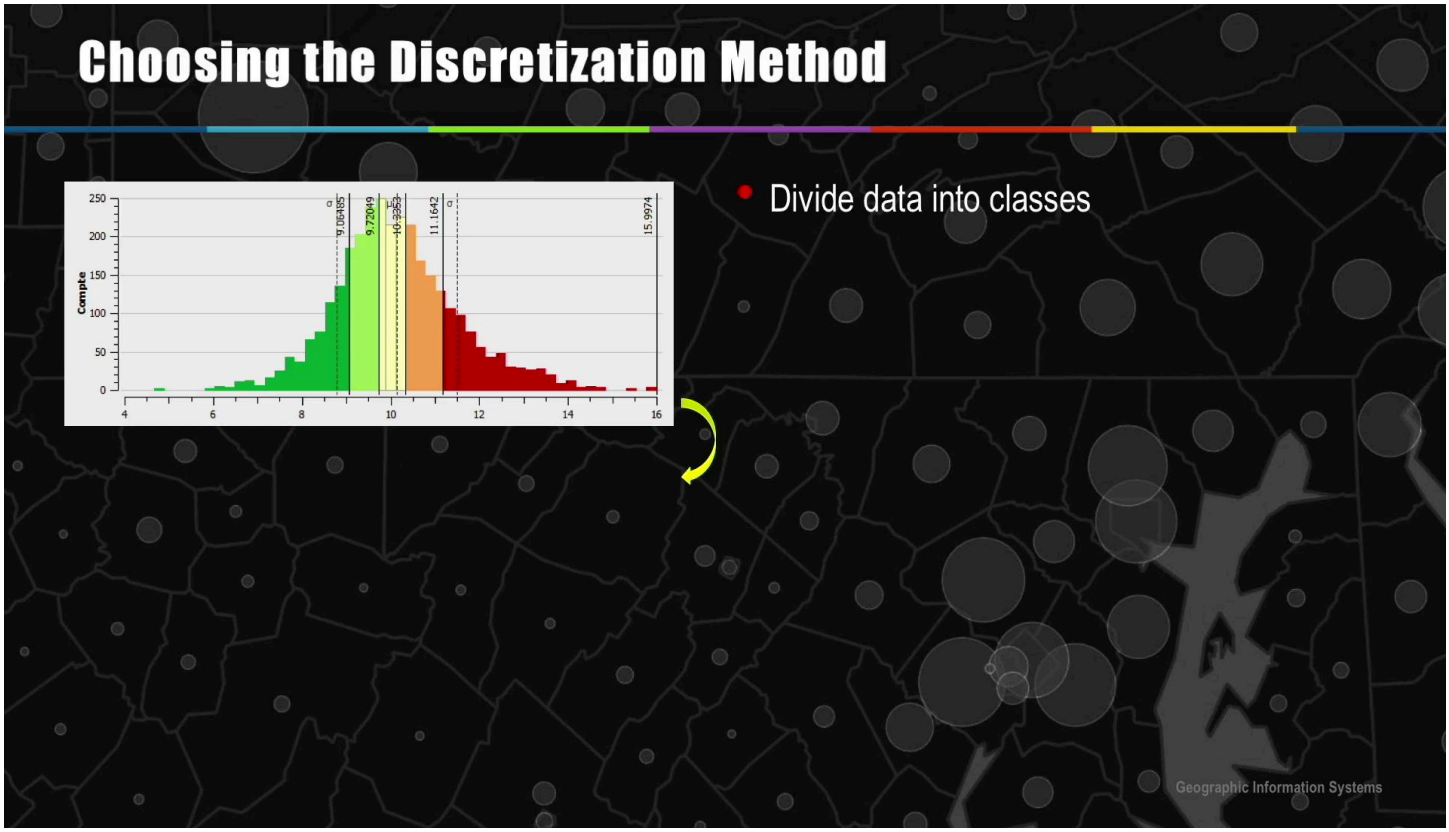
Geographic Information Systems

On this map of american counties in seven classes and two shades, the dark green represents a low rate, so a favorable situation. The value of the green decreases towards the intermediate class in pale yellow, then the red value increases towards the highest homicide rates in counties where the situation is negative. We exploit here the meaning associated with the green and red colors by analogy with the highway code and the traffic lights. The explanations are all the more important that this subterfuge induces the creation of a theoretically false map since the statistical distribution of a variable such as the homicide rate should be represented by a simple gradient of valuesof a single shade.

Notes

Summary

26m 25s

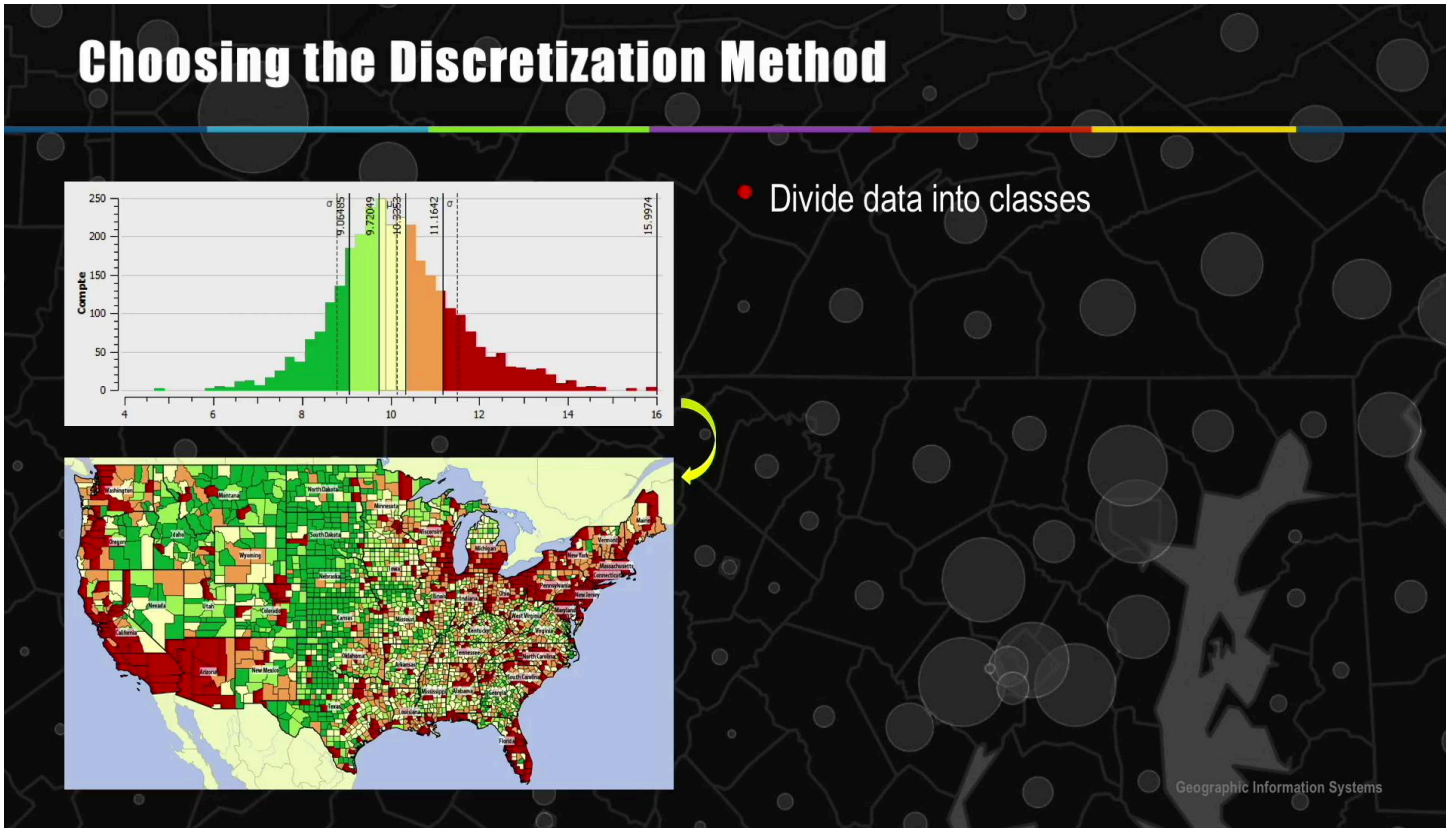


Once we have an idea of the number of classes to use, we have to distribute the data of the statistical distribution to be represented.

Notes

Summary

27m 11s

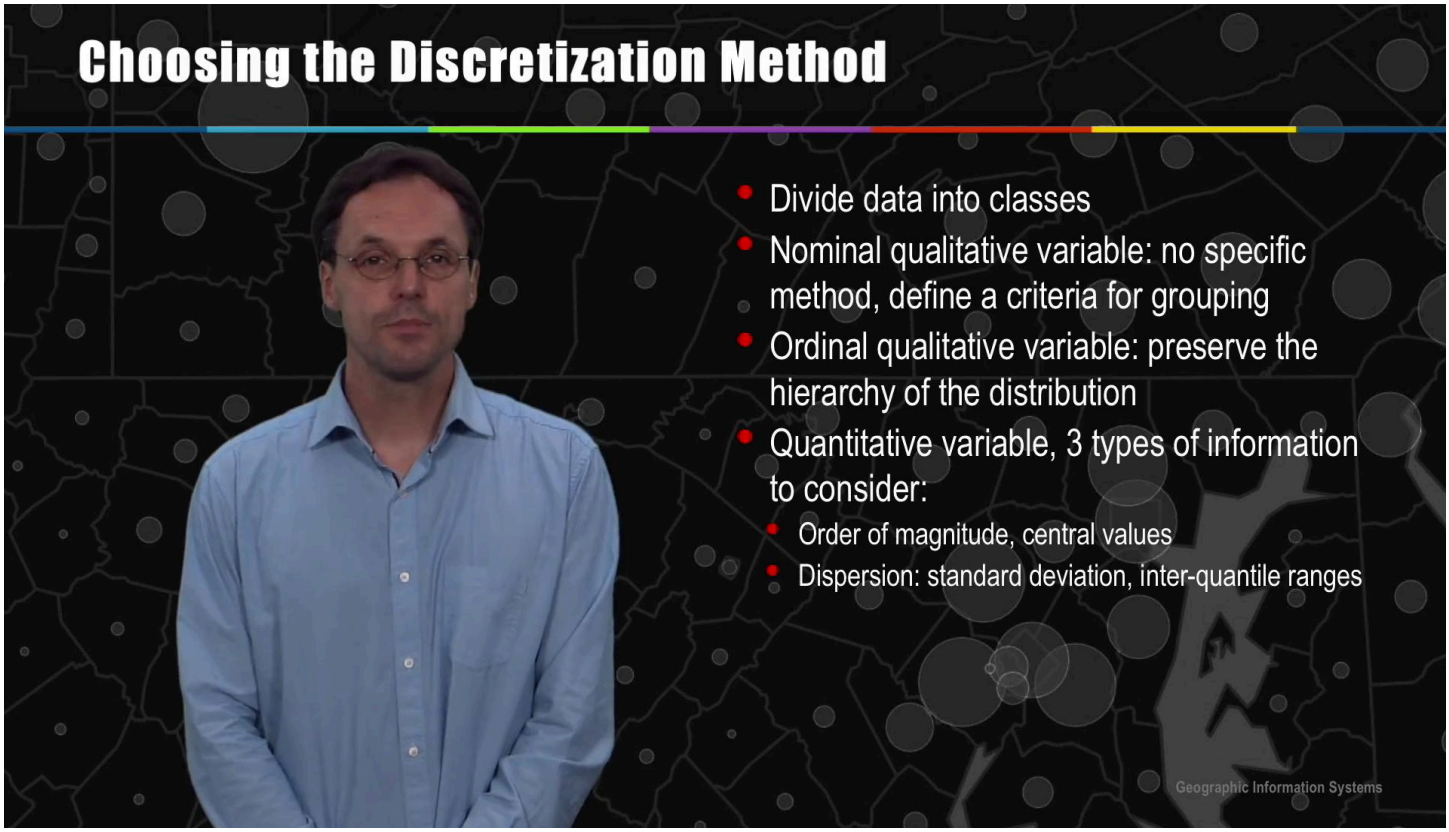


The principles of the discretization differ depending on the nature of the information.

Notes

Summary

27m 20s

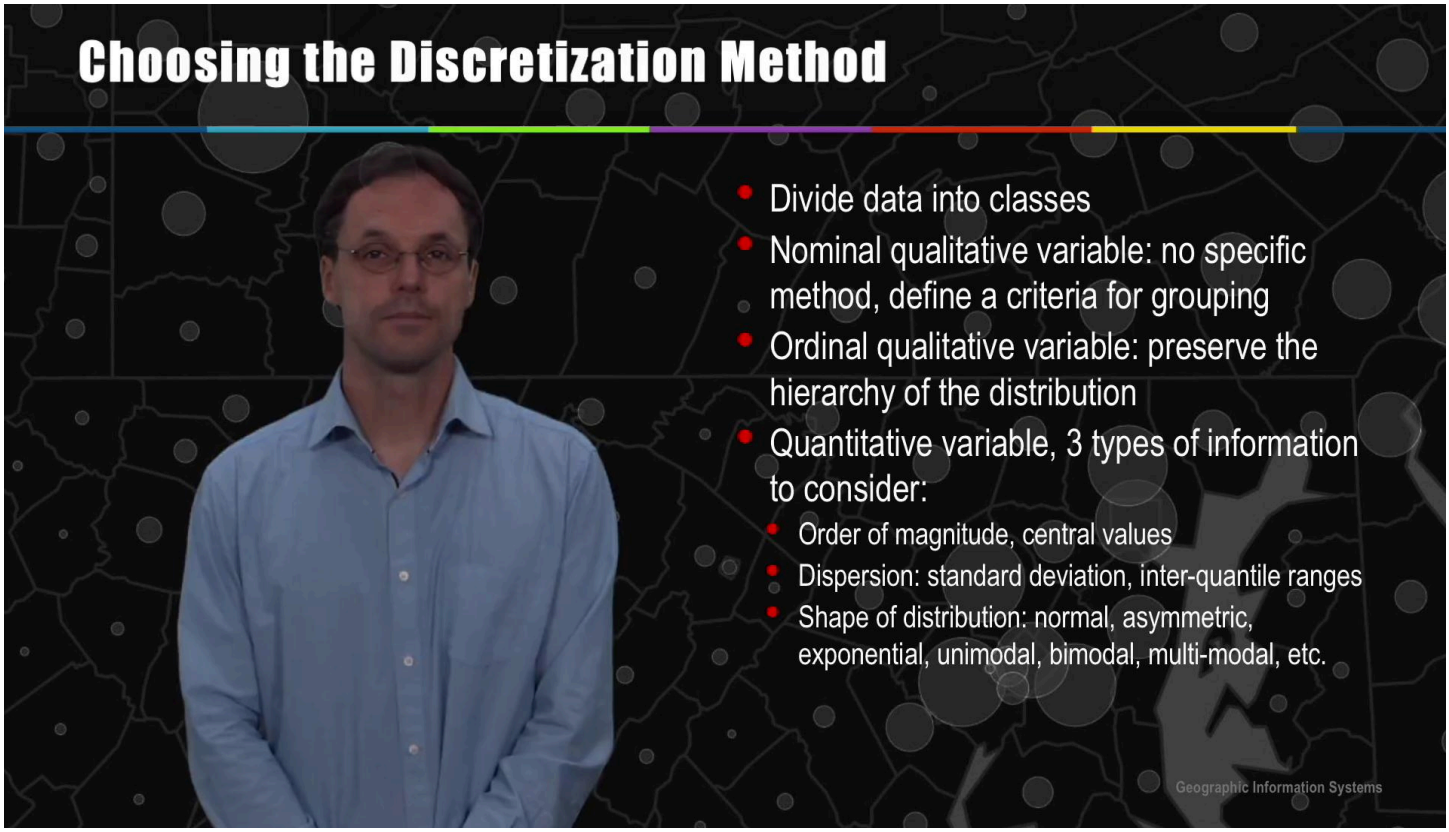


If the information is nominal qualitative, as we have seen, there is no specific method, we will seek to define a common criterion of grouping to get to the construction of a typology. Each grouping of geographical objects is specific to simplification objectives of the selected data. This is the case of the map of american counties grouped according to their belonging to the states which we developed in QGIS a little while ago. If the variable is ordinal qualitative, we will try to preserve the hierarchy of information and to constitute the classes on the basis of the ranks, as we also discussed earlier. In the case of quantitative information, three types of distribution information can be taken into account. These information will help determine which method of discretization to use. The first type is the order of magnitude which is revealed by the central values of the distribution. This is the mode, the mean or the median which can be selected as class limits. Then, the measurement of the dispersion is a measure of the inequality of values, it is characterized by the standard deviation or by an inter-quantile interval, values that can be selected as amplitudes of classes.

Notes

Summary

27m 24s



A measure of dispersion takes into account the variance and enables to minimize the differences between objects of the same class and to maximize the differences between classes. Finally, the shape of the distribution informs about the behavior of a phenomenon and it has to be taken into account to determine the most appropriate class setting method.

Notes

Summary

28m 36s

Discretization Methods

Method	Definition	Calculation	Comments	Type of Distribution
Standard deviation	All classes, except for the two extremes, have the same range that is equal to the standard deviation	Mean and standard deviation	If there is an odd number of classes, the middle class rests on the mean. Interest: dividing classes with respect to the mean highlights extremes and enables comparison between maps	Normal (Gaussian), with a concentration of data around the mean, may be slightly asymmetric
Equal interval	Constant and equal intervals	(Maximum value – minimum value) / number of classes	Simple method that is easy to use and interpret, but seldom used because it is not appropriate for skewed datasets: classes can be very different, some could be empty. Maps cannot be compared.	Uniform distribution (data are uniformly distributed across all possible values), normal (Gaussian) with a concentration of data around the mean.

Source: based on "Rappels sur les discrétisations», F. Demoraes, M. Souris, T. Serrano, Laboratoire de Cartographie Appliquée, Paris X. Nanterre, & E. Ployon, Université de Savoie, Formation SIG-Santé 2010

Geographic Information Systems

We present here five methods with different characteristics. The mean deviation method Allows to generate classes of identical range, except the extreme classes which are different. The range of a class is equal to the standard deviation. In the case where the number of classes is odd, the middle class will be astride the average. The main interest of this method is that it makes the comparison possible between several maps, but the distribution of the analyzed variable must be normal or weakly asymmetrical. The method of equal intervals or classes of equal amplitude aims to constitute classes whose intervals are constant. We calculate these intervals by subtracting the minimum value from the distribution to the maximum value and by dividing the result by the number of classes previously determined. This method is simple but not often used because the risk is to fall on empty classes or very unequal numbers. In addition, it does not allow comparisons between maps. This method requires a normal and uniform distribution, which is very demanding.

Notes

Summary

28m 57s

Discretization Methods

Method	Definition	Calculation	Comments	Type of Distribution
Natural breaks (jenks)	Observed thresholds	Look for natural breaks in the sorted frequency histogram	Takes discontinuities in the dataset into account. Maps are not comparable	All distributions with discontinuities Multi-modal distributions
Quantiles	Each class has the same number of units in it	Total datapoints / number of classes Class limits = nb of individuals, ordered statistic	Does not take extreme values into account. The chosen limits could be debatable (similar values divided into adjacent classes) Allows for comparisons	Uniform distribution, or other distribution without discontinuiuties Should be avoided if there are extreme values or a large number of similar values.
Box map	Classified according to boxplot or box and whisker plot	6 classes: quartile 1 (Q1), Q2, Q3, Q4, lower outliers (Q1 – [1.5*(Q3-Q1]), upper outliers (Q1 + [1.5*(Q3-Q1])	Ordered statistic, allows for comparisons, can be implemented in Geoda.	All distributions, can be used if there are extreme values.

Source: based on "Rappels sur les discrétisations», F. Demoraes, M. Souris, T. Serrano, Laboratoire de Cartographie Appliquée, Paris X. Nanterre, & E. Ployon, Université de Savoie, Formation SIG-Santé 2010

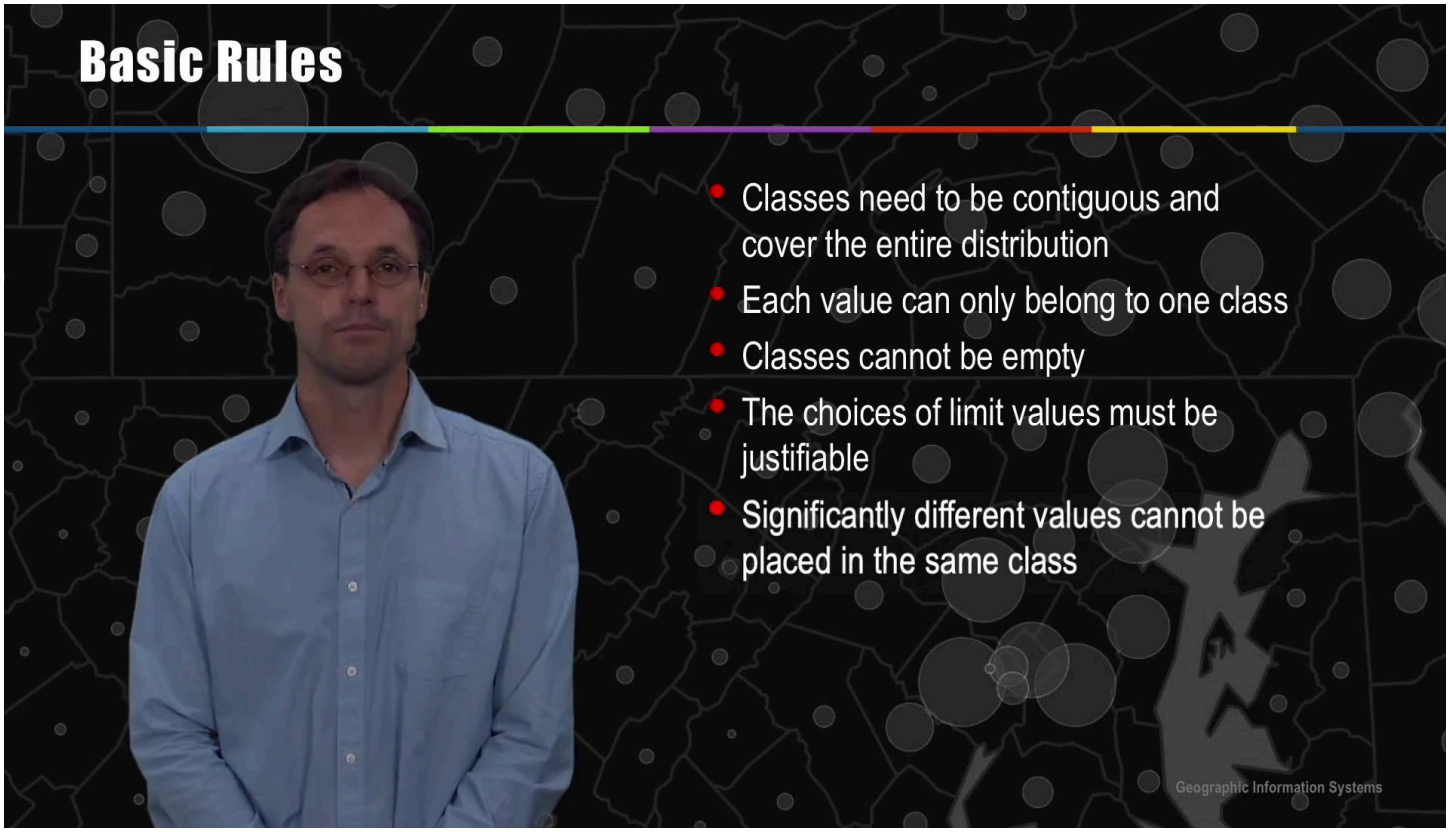
Geographic Information Systems

The method of natural thresholds is based on the observation of the histogram of cumulative frequencies sorted in ascending order. If they let appear net thresholds between groups of spatial units, these are used as a class limits. This method exploits the discontinuities of a distribution, but does not allow the maps to be compared with one another. This method is suitable for multi-modal distributions. In the case of the quantile method, each class has the same number of individuals. It is a statistic of order and the number of classes is defined by the total number of individuals divided by the number of classes. One default of this approach is that it cannot take outliers into account and certain class limits can be debatable since very close values can fall into distinct classes. On the other hand, this type of class setting allows to make comparisons between maps, to be avoided if the distribution presents extreme values. Finally, the box whisker method also exploits a statistic of order. Its principle is to take the classes generated by the box-plot, an exploratory analysis tool developed by John Tukey. It produces six classes four of which are the quartiles, and two are classes allowing to highlight the extreme valueson both sides of the distribution. The maps produced with the help of this method are comparable.

Notes

Summary

30m 06s

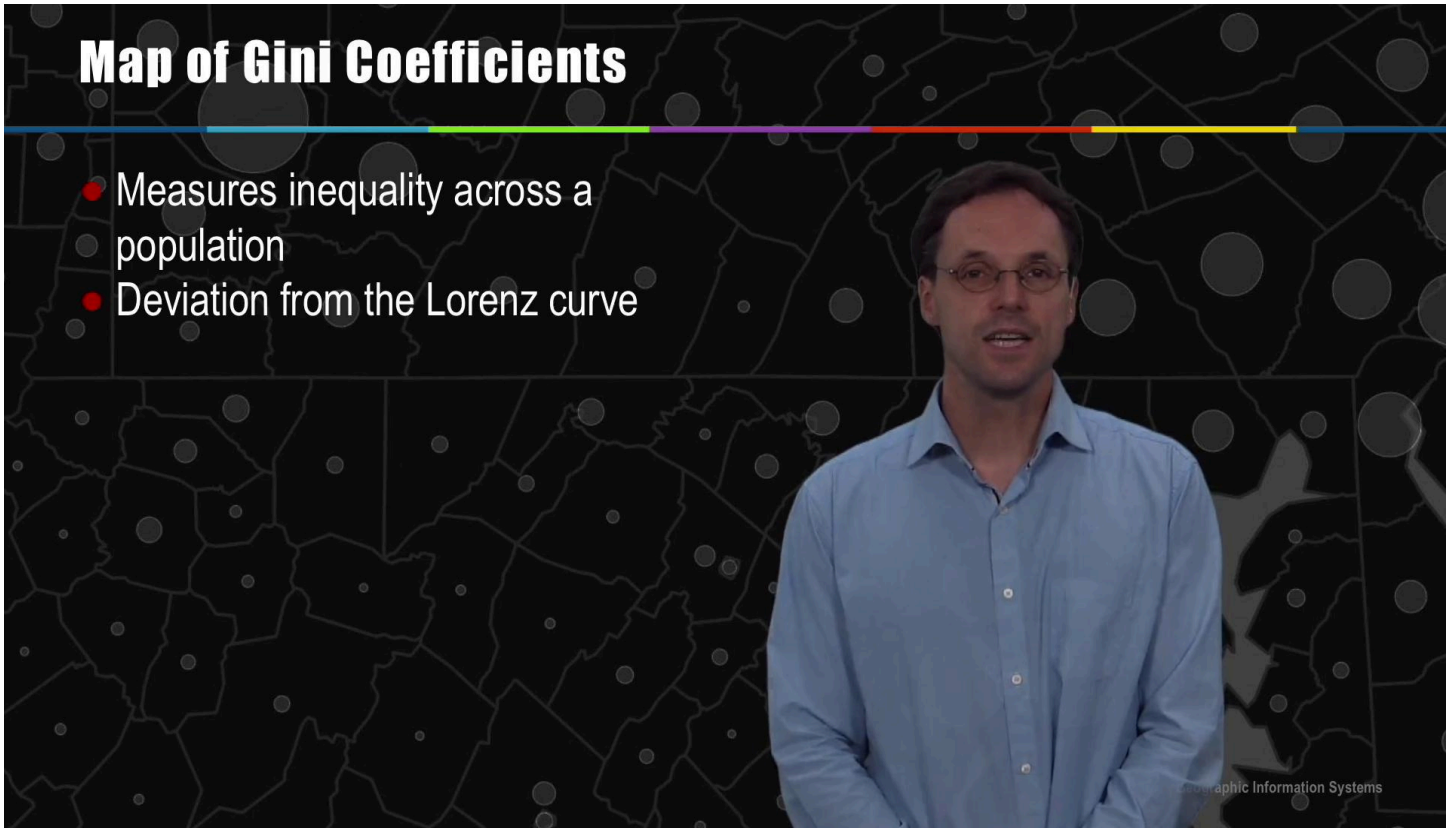


Finally, by applying one or the other method of class setting, we have to make sure that we respect the following rules: the classes must cover the whole statistical distribution and must be contiguous. Any value can belong to only one class. No class can be empty. The choice of limit values must be based on clear criteria. And the values that are very close should not be placed in different classes. We can now see how to create a choropleth chart.

Notes

Summary

31m 34s



The aim is to analyze the spatial distribution of the Gini coefficient in the american counties in 1989. This coefficient is used to measure the inequality in incomes within the populations of the different counties. This level of inequality can then be compared from one spatial unit to the other. The Gini coefficient is a number that varies between 0 and 1 where 0 means a perfect equality and 1 a total inequality. The Gini coefficient is calculated in relation to the function which associates with each fraction of the population getting an income ranked in ascending order, the share that these incomes represent.

Notes

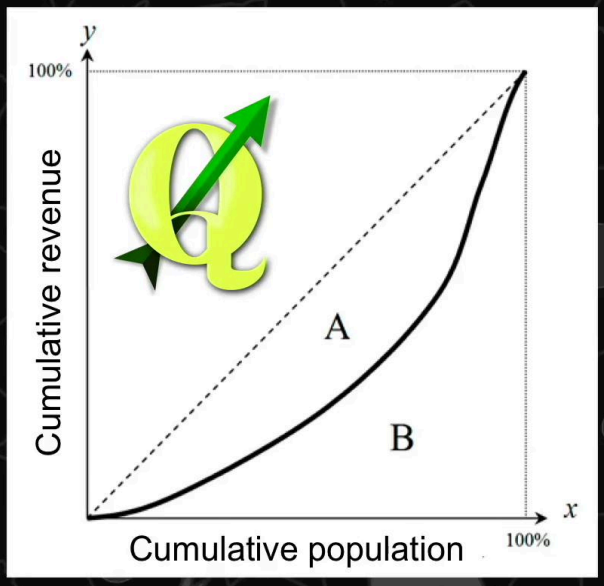
Summary

32m 10s

Map of Gini Coefficients

- Measures inequality across a population
- Deviation from the Lorenz curve
- Further the curve is from the bisector, the greater the inequalities
- Coefficient = *A/(A+B) = A/(1/2) = 2A = 2(A+B-B) = 2(A+B) - 2B = 1 - 2B*

C. Gini (1921) Measurement of inequality of income, Economic Journal 31, 22-43.



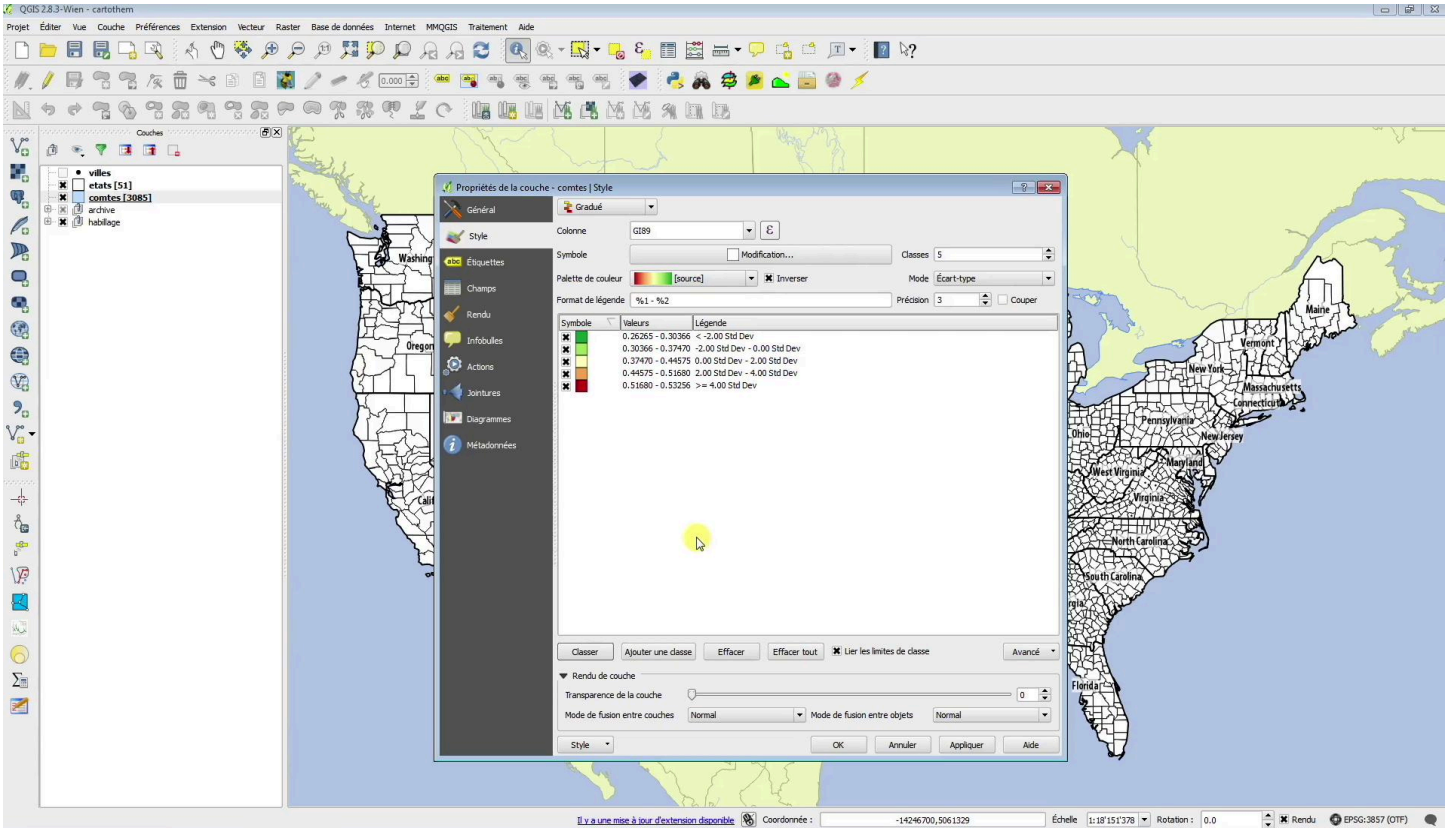
Geographic Information Systems

The Gini coefficient estimates the inequality by the deviation from the Lorenz curve which illustrates the distribution of wealth in a society. The farther this curve is from the bisector, the greater the inequalities. The Gini coefficient is the area difference between the triangle formed by the bisector and the area delimited by the income curve. For more details, we refer you to the original article from Corrado Gini. We can now look in QGIS.

Notes

Summary

32m 50s

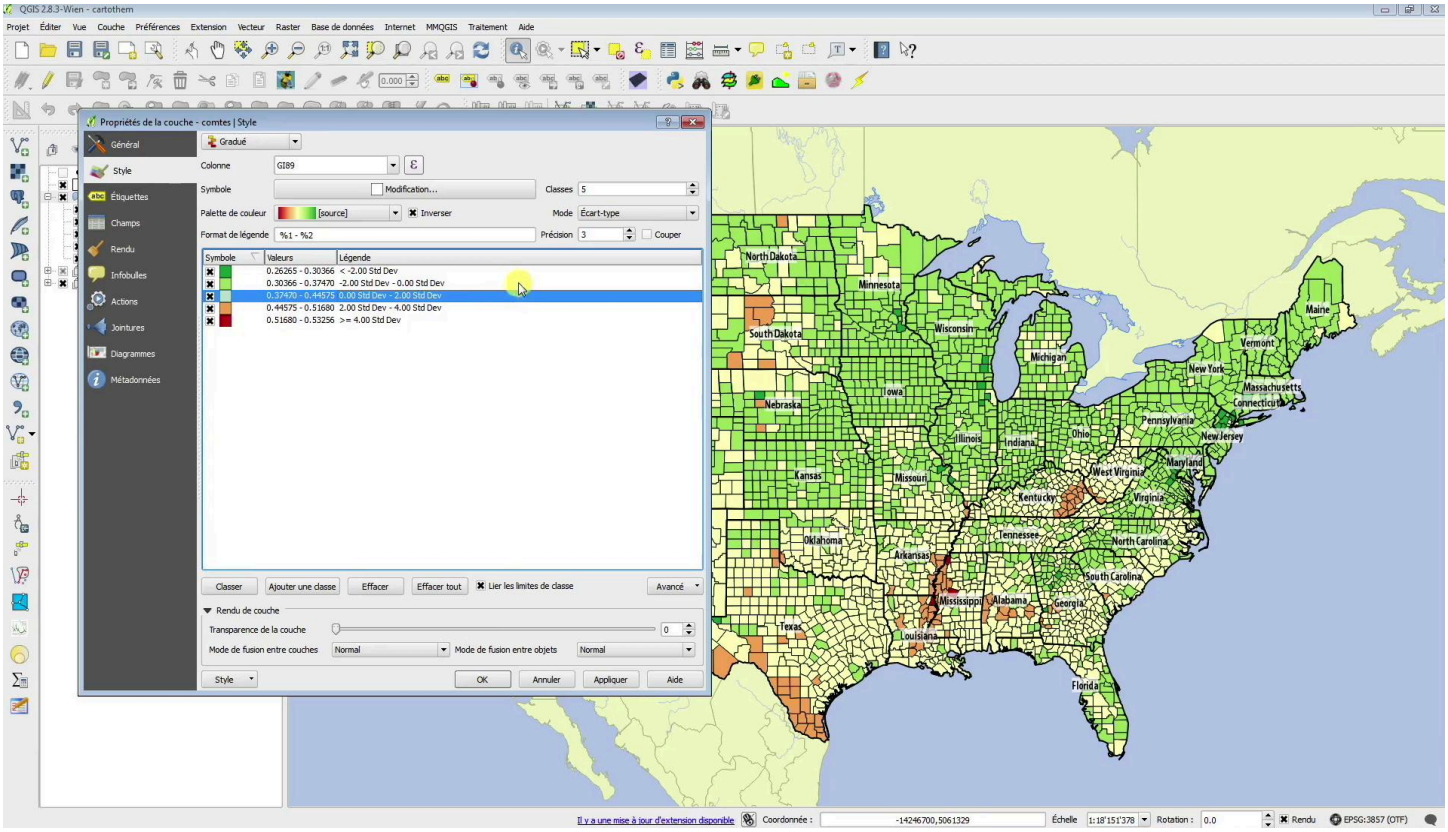


First, we will take a look at the statistical distribution of the Gini coefficient variable of american counties for 1989 and called GI 89. The form of the distribution will help us determine the appropriate method of discretization. In QGIS, a project has already been opened containing the layer of american counties. In the "Statist" extension, which you will have previously installed, the "county" layer is selected, then the variable "GI 89" before clicking on OK. The distribution is almost normal, slightly asymmetrical, with an average a little bit higher than the median. In this case, we can use the mean deviation method which allows to generate classes of identical range with the exception of extreme classes. The extent of a class will be equal to the standard deviation. To do so a double-click on the "county" layer allows to reach the properties of the layer and the "style" tab. In the drop-down menu located at the top of the window, choose "graduate", then select the "GI 89" column. Choose five classes and the predefined color palette "red, yellow, green" which is divergent with a central class in light yellow and then we will reverse the palette so that indices of high inequality are displayed in red then, we choose the "standard deviation" discretization mode and finally, we click on the "sort" button.

Notes

Summary

33m 19s

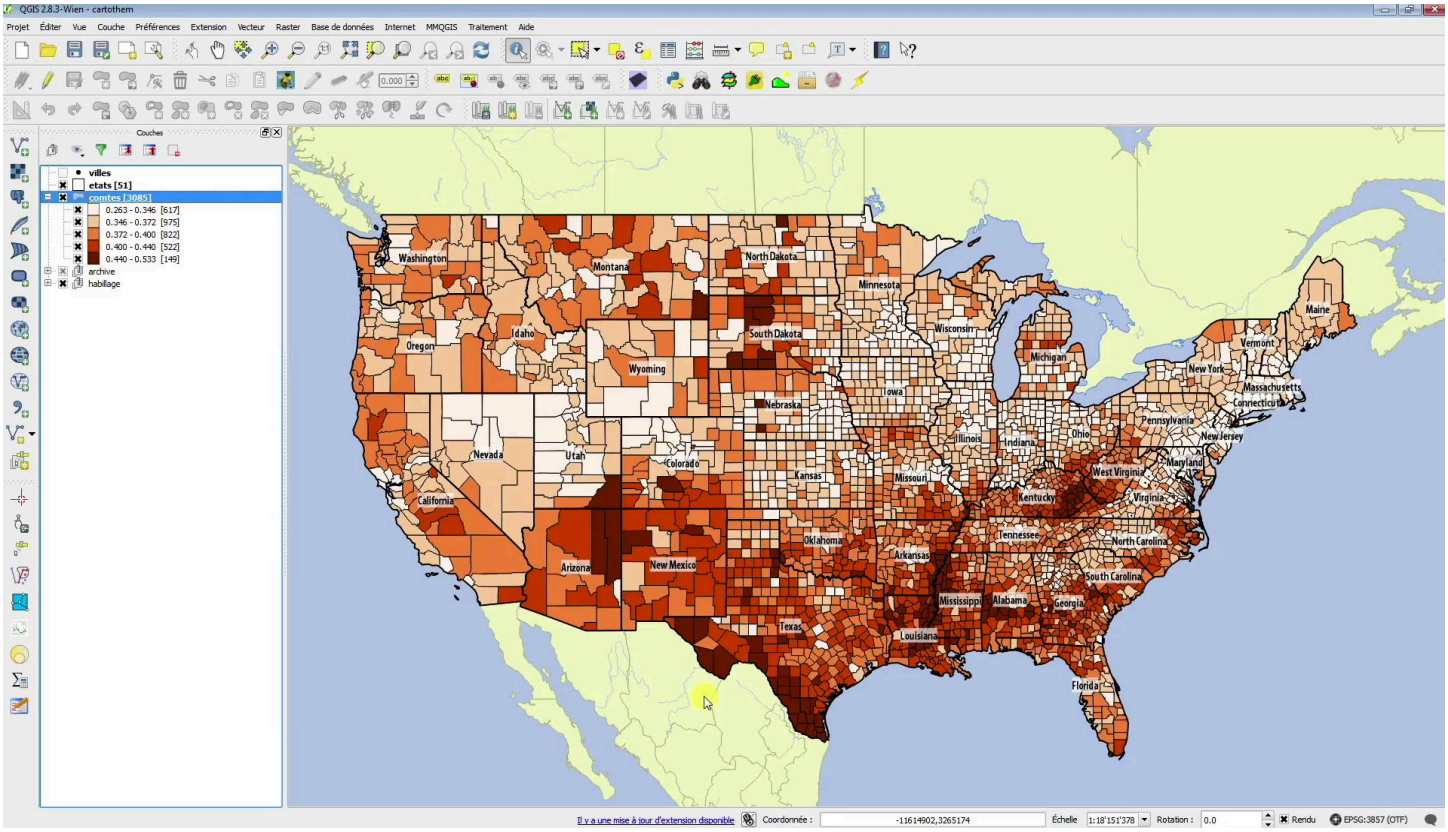


At this stage, it is possible to use the "ColorBrewer" tool described in the previous lesson. We can indeed use a carefully developed palette in order to optimize the perception of thresholds between the shades, although the basic palettes provided by QGIS most often provide an acceptable solution. On the ColorBrewer interface, we have to choose the desired parameters, so the number of classes, here five, the type of data, here diverging, then the corresponding "red, yellow, green" palette. Depending on the aim, it is also possible to specify certain options including the generation of palettes usable by people with perception anomalies. And then, we have to transfer the values of the classes in hexadecimal format for example, by copying and pasting between ColorBrewer and QGIS. Once the shades have all been assigned, we can click on OK and we obtain the following result. You will notice that for this mode of discretization based on the mean deviation, QGIS permits only five or nine classes, which implies that the class between 0 and 2 standard deviation takes the neutral pale yellow shade, while the symmetric class between 0 and minus 2 standard deviation takes a light green, non-neutral tint.

Notes

Summary

34m 56s



In this case, we should be able to resort to an even number of classes and a palette in two shades, red and green, so as to be able to represent the opposition of behavior between these two classes on the map. Nevertheless, the map produced clearly shows the southern Texas counties on the border with Mexico, and along the Mississippi River in the South of the United States. We can also sequentially use the distribution of Gini's values944 00:36:42,267 --> 00:36:44,960 with a gradient of the orange shade by applying the natural threshold method in five classes. This representation better highlights the counties where the inequalities are the strongest.

Notes

Summary

36m 16s