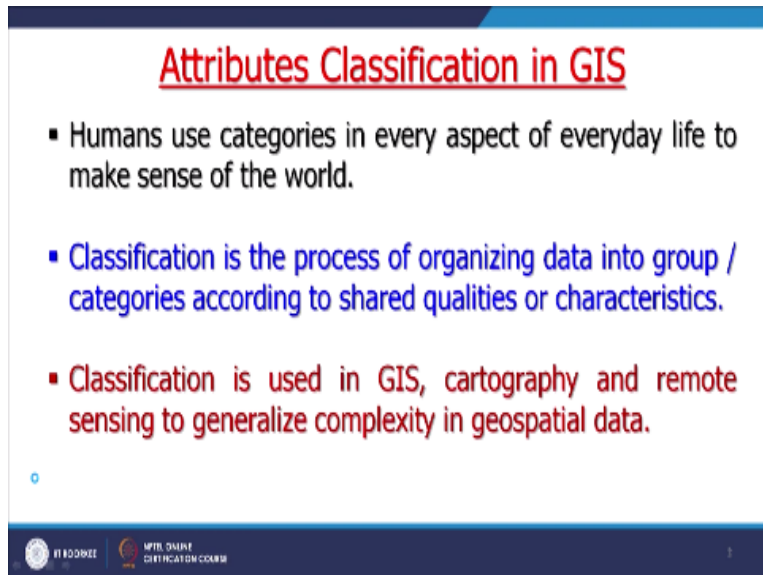


Geographic Information Systems
Prof. A. K. Saraf
Department of Earth Sciences
Indian Institute of Technology - Roorkee

Lecture - 28
Attributes Classification Methods

Hello everyone! and welcome to a new discussion which we are going to have one attributes classification methods. As you know that our GIS data is of 2 types. One is spatial data which includes vector data, raster data and TIN and various types of vector data, 2 types of raster data and one type of TIN. Same way, we also have the non-spatial data or attributes data. And we will be going through the classification methods, how these can be classified? And we can take advantage of different classification methods which are available.

(Refer Slide Time: 01:04)



Attributes Classification in GIS

- Humans use categories in every aspect of everyday life to make sense of the world.
- Classification is the process of organizing data into group / categories according to shared qualities or characteristics.
- Classification is used in GIS, cartography and remote sensing to generalize complexity in geospatial data.

IIIT Roorkee IIT Roorkee
M.Tech. Degree CERTIFICATION COURSE

So, in the classification basically when we are having a continuous data, sometimes it is difficult to make sense out of continuous data. And therefore, our brain or humans are more comfortable to use data in different categories, means in different classes. And therefore, when we are having a continuous data, sometimes we have to do you know sort of discretization.

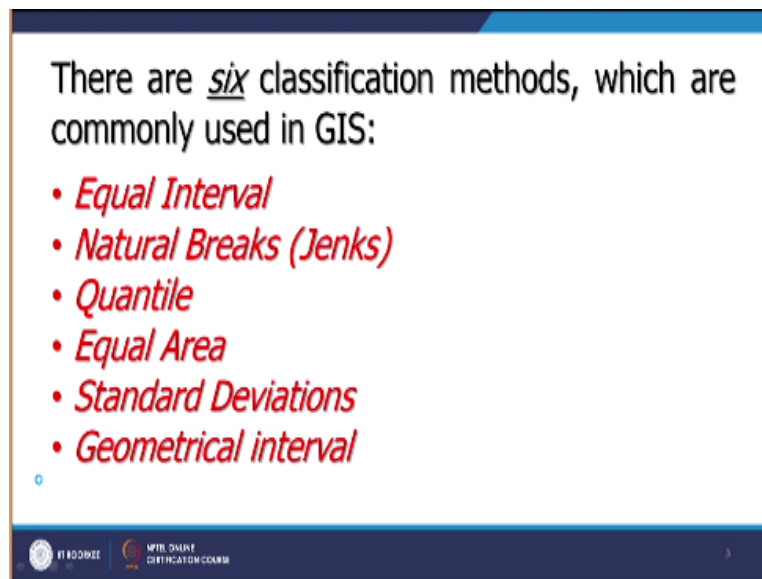
For example, if I am having a digital elevation model which will be showing a continuous elevation values for a terrain of area but when if I want to create a classification that means using certain classification, I can create a relief map which may be more understandable to

many humans. So, use of categories is very common in that. Now basically, what we do in classification that we organise or reorganise data into different groups or categories.

And sometimes maybe based on statistical techniques or sometimes as per our requirements also that we can group them into or categorize them according to our requirements or depending on the characteristics of the attribute data as well. Now, classification in GIS is used in cartography, in remote sensing. And in the field of digital image processing of remote sensing, there are lot of classification techniques are there, even some are based on neural network and most modern techniques are there.

But in GIS especially for attribute classification, we not need to go to that extent. Very simple classification methods are available. And so far, the 6 classification methods for attribute data classification have been implemented into GIS which we see commonly in different GIS softwares. So, we will take one by one in this discussion. But let me first bring all 6 classification methods.

(Refer Slide Time: 03:18)



So, first one which is the most common one is the equal interval. And by default, also in many softwares, you would find that whenever you try to attempt a classification of attribute data, this will you would see in the default, equal interval; dividing. For example, you are having a digital elevation model, values are varying between say 0 to 100 metres.

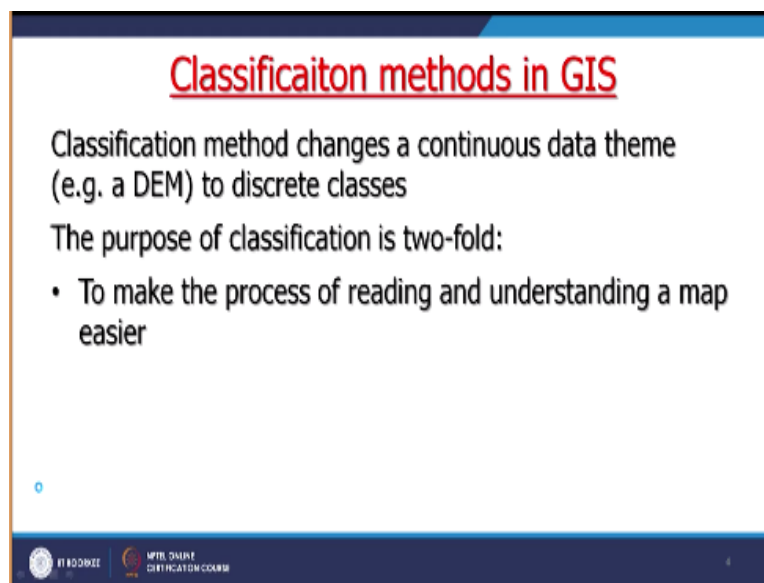
Now, you can classify in say 5 categories; 0 to 20, 21 to 40, 41 to 60, something like that. So, dividing in equal slices like a loaf of a bread. Now, there can be another method which is

called natural breaks or Jenks which is based on Jenks algorithm. And this is very common. You know we also use in case of awarding grades to the students. We find the natural breaks. Where large gap is there, a new class can be created something like that.

So, a new grade can be assigned. So, it is based on again statistical technique and especially given by the Jenks or Jenks formula. Third category or third type of classification method is quantile which we will see in details. Equal area or rather than having controls over you know attribute values, we can involve even area that I want equal area classification. Now thus, a number of equal pixels or cells will also be counted in this classification which we will see again detail about this.

And then fifth one is the standard deviation; typical statistical technique. And the most recent one which has been implemented in various GIS software is geometrical interval.

(Refer Slide Time: 05:13)



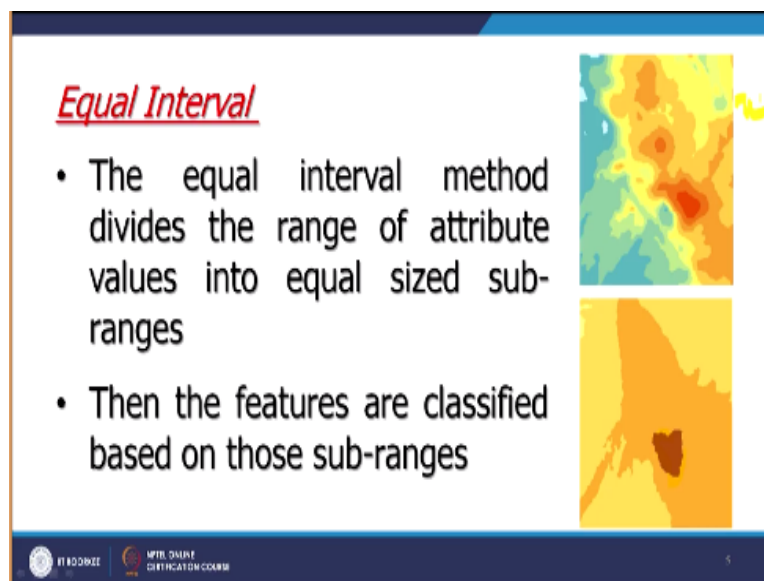
So, let us start with the first method which is the classification based on equal distance. And while we are doing as we know that from continuous data, we want to discretize or bring them in different categories. And as you know the purpose of these classification techniques are basically 2-fold to make reading and understanding a map or a digital elevation model or raster data much easier.

For example, if you are having a colour composite of a satellite image. Now, if some expert can classify that image and can provide a map say, land use map of different categories, that

would be more useful for many people rather than just going through the satellite image. Because many people may not be able to interpret satellite images easily as an expert.

So, an expert can do the classification into different categories and then that map becomes much more understandable and usable also. And also, sometimes we use classification to hide something or to highlight something. And that way, it is also very-2 useful. Again, in digital elevation or in digital image processing techniques, there is a technique which is called density slicing or masking so we can also do that kind of classification.

(Refer Slide Time: 06:49)



Equal Interval

- The equal interval method divides the range of attribute values into equal sized sub-ranges
- Then the features are classified based on those sub-ranges

The slide contains two maps of a geographical region. The top map is a digital elevation model (DEM) showing a continuous range of elevations with a color gradient from blue (low) to red (high). The bottom map shows the same region after classification using the equal interval method, resulting in three distinct color categories: light yellow, orange, and dark brown.

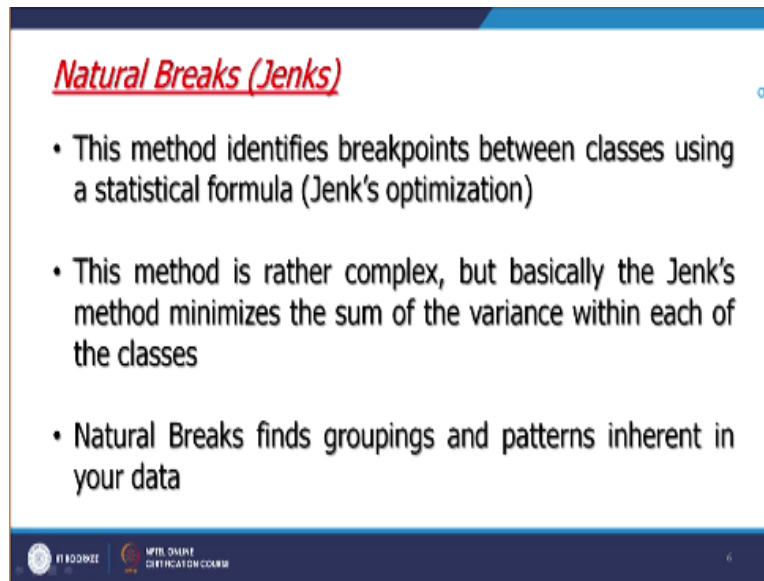
At the bottom of the slide, there are logos for 'Uttarakhand State Open University' and 'M.Tech. Degree Certification Course'.

Now, first one is equal interval. How equal interval is very simple that the entire range of attribute values are divided into equal size categories or sub ranges. Here an example is given. At the top, what you are seeing a digital elevation model. And then it has been categorised into 3 equal sized sub categories or categories having different elevation ranges.

And then though, if I am interested in particular height or you know elevations, for some kind of project then I can focus very well in the centre part rather than maybe little difficult for this continuous data. So, this is basically what we are doing. A continuous data is being discretized into different categories. So, this category word is also used. Instead of directly uses the classification, software like ArcGIS; they use the word category.

So, you may not be getting directly the word this one but when you go in the category then you get the different classification method. And we will have demo also on these classification methods as well.

(Refer Slide Time: 08:11)



Natural Breaks (Jenks)

- This method identifies breakpoints between classes using a statistical formula (Jenks optimization)
- This method is rather complex, but basically the Jenks method minimizes the sum of the variance within each of the classes
- Natural Breaks finds groupings and patterns inherent in your data

WPI BOSTON WPI ONLINE CERTIFICATION COURSE

Now, the second one is the natural breaks which is based on the Jenks optimization or Jenks formula. And basically, this method identifies the break points or a sufficient or large gap in the data which is based on a statistical formula. Now, this method is of course, complex than equal interval method relative to that. But basically, what this method does? It minimises the sum of variance.

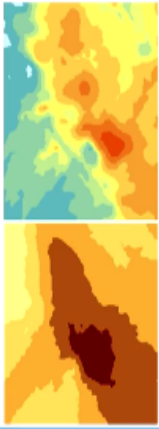
So, within a category, the variance is very little and with between 2 categories, you are having a large gap. So, this is what the Jenks optimization technique is that it minimises the sum of variance within each of the class. And natural breaks find basically groupings and patterns inherent in your data which you may not identify; may not see or may not be visible to you in a continuous data.

But when you know put this type of classification over that data then you can find the natural breaks, groupings and of course, whatever the pattern which was inherent in the data set will also get highlighted.

(Refer Slide Time: 09:26)

Quantile

- In this method, each class contains the same number of features



UIN AR-RANIRI IPTK SURABAYA CERTIFICATION COURSE

Third type of attribute classification method is quantiles. Again, in quantile; the same digital elevation model is taken. And in the next, the output; you are having different classes and these classes are having basically say, frequency-based classification. So, that means each class is having same number of cells or pixels in it. And these quantile classes are perhaps the easiest to understand.

But they can be sometimes misleading from the input data point of view. So, one has to be really careful and remember only that first rule of GIS that after each and every step, check for errors. So, if it is misleading, you may choose some other method of attribute classification. Now here because it is frequency based so the total count is there; the population count and which is just opposed to the density or percentage.

(Refer Slide Time: 10:36)

- Population counts (as opposed to density or percentage), for example, are usually not suitable for quantile classification because only a few places are highly populated.
- One can overcome this distortion by increasing the number of classes.
- Quantiles are best suited for data that is linearly distributed

UIN AR-RANIRI IPTK SURABAYA CERTIFICATION COURSE

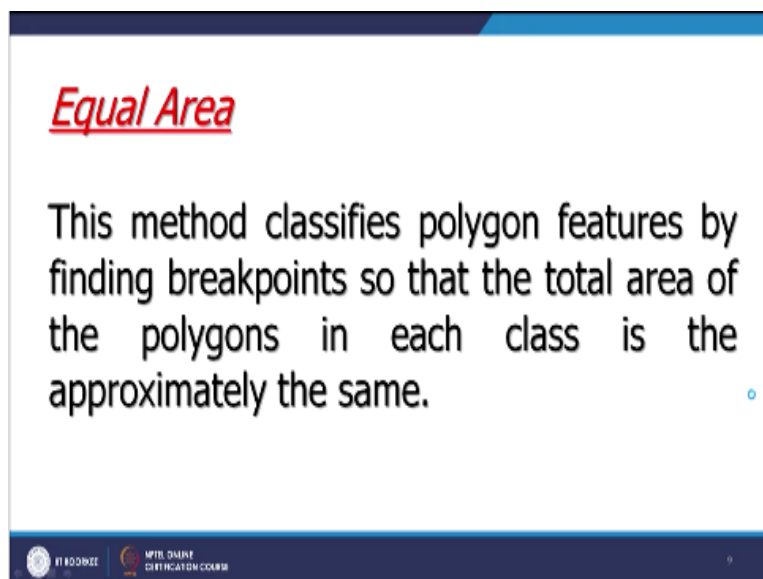
For example, population count is usually not suitable for this quantile classification. Only few places are highly populated. So, if you are having a surface which is representing population counts and which is already having some concentration of population or high density of population then this kind of classification may not be good. What it means also that all classifications are not suitable for all types of data.

So, it depends on how the data is distributed within your data set or within one file. If data is well distributed then various types of classification can be employed. But if you know like already inherent groups are there then all classification will not suit that one. So, one has to be a little judicious about that. And before that, the best technique is to study the histogram or distribution of the values with reference to the frequency through this histogram.

And study of histogram will let you know that which classification, you can adopt and can give better results rather than just trying one after another. So, always it is better to study simple statistic like what is the minimum, maximum, mean, mode and then histogram. And the histogram; when we will have the demo, we will discuss that part also that how after seeing the histogram, you can decide.

And if you are having highly populated places within one data set then a small number of classes may give you wrong interpretation of the data set through classification. So, the best thing is in that case that you can increase the number of classes. Now, if data is linearly distributed then quantile is the best suited classification for attribute data.

(Refer Slide Time: 12:51)



Equal Area

This method classifies polygon features by finding breakpoints so that the total area of the polygons in each class is the approximately the same.

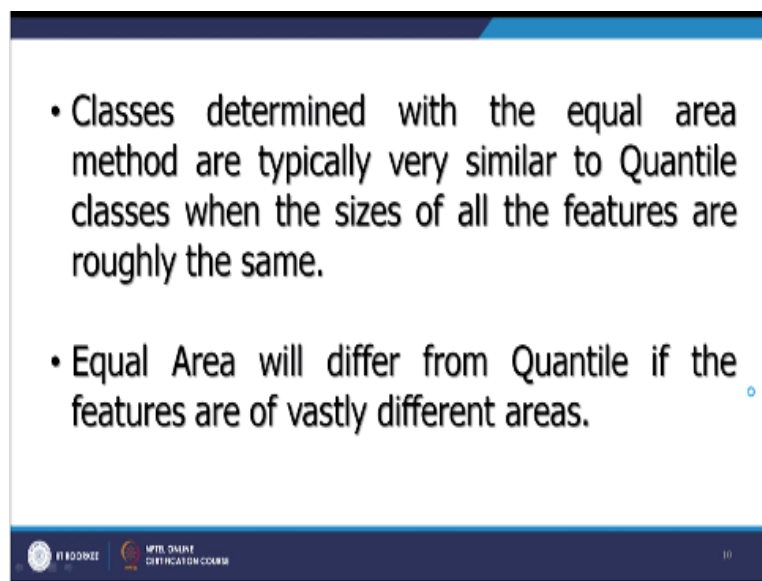
9

UPTU ONLINE CERTIFICATION COURSE

Now, this one is the equal area one. In this fourth method, the equal area says that now it is going to again count the number of cells because each cell will have its own spatial resolution so by which, we can bring the area also. So, this method basically classifies as a polygon feature. Also, you can apply on the raster data by finding breakpoints so that the total area of the polygons in each class is approximately the same. That is why, it is called equal area.

So, in some special kind of requirements or projects, we may resort to this type of classification also.

(Refer Slide Time: 13:45)



Now here in equal area classification, the classes are determined with equal area method which are similar to the quantile, when the sizes of all features are roughly the same. Because in quantile also, we are using the frequency of the occurrence of individual cell values. Now, equal area will differ from quantile on one account that features are vastly different areas.

(Refer Slide Time: 14:06)

Standard Deviations

In this method, the mean value is found and then class breaks above and below the mean at intervals of either $1/4$, $1/2$, or 1 standard deviation are placed until all the data values are contained within the classes.

Now, the fifth one is the standard deviation which is purely a statistical based technique. So, in this method, the mean value is found first and then class breaks above and below mean at intervals of $1/4$, $1/2$ or 1 standard deviation are placed until all data values are contained within the class. This is a very standard way of you know using standard deviation method.

So, by this, we can know that you know different classes and which are below standard deviation, which are higher standard deviation and the mean part also. As said also that before going for classification, it is always a good to first check simple statistics; minimum, maximum, mean, mode and histogram. And then you can know very well that which classification should be employed.

(Refer Slide Time: 15:10)

Further, values are aggregated those are beyond three standard deviations from the mean into two classes, greater than three standard deviations above the mean (" > 3 Std Dev.") and less than three standard deviations below the mean (" < -3 Std. Dev.").

Now in this standard deviation classification, the values are aggregated and those are beyond 3 standard deviations from the mean into 2 classes; > 3 standard deviations above mean and < -3 standard deviation below the mean. So likewise, $-3, +3$; all these values can be accommodated. After this discussion, we will also see a comparison of different classifications over the same data set.

And we will see that how different classification can bring completely different results on the same data set. So, before that, I would like to discuss this last classification method which has recently been employed or implemented into GIS that is geometrical method.

(Refer Slide Time: 15:55)

Geometrical interval

- This is a classification scheme where the class breaks are based on class intervals that have a geometrical series.

IT KOOBIZ | NTEL ONLINE CERTIFICATION COURSE

Now these are like from quantile. There is slight variation to equal area. Similarly, it is also slight variation. So, these you know variants now have started coming. In this geometrical classification method where the class breaks are placed on class intervals that have a geometrical series. So, using that geometrical series, we can classify. So here on the left data or left image, what you are seeing the original input digital elevation model and through this geometrical classification when we perform, we get a result something like that.

So, different classification methods will produce different results. If anybody is interested further to know about the geometric series and how it has been implemented then here you can find detail even in the Wikipedia.

(Refer Slide Time: 16:52)

Geometric series

From Wikipedia, the free encyclopedia

This article is about infinite geometric series. For finite sums, see geometric progression.

In mathematics, a **geometric series** is a series with a constant ratio between successive terms. For example, the series

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \dots$$

is geometric, because each successive term can be obtained by multiplying the previous term by $\frac{1}{2}$.

Geometric series are among the simplest examples of infinite series with finite sums, although not all of them have this property. Historically, geometric series played an important role in the early development of calculus, and they continue to be central in the study of convergence of series. Geometric series are used throughout mathematics, and they have important applications in physics, engineering, biology, economics, computer science, queueing theory, and finance.

Each of the purple squares has $\frac{1}{4}$ of the area of the next larger square ($\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$, $\frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$, etc.). The sum of the areas of the purple squares is one third of the area of the large square.

Contents [hide]

- Common ratio

IT FORGE RMIT ONLINE CERTIFICATION COURSE 11

And this is of course based on mathematical geometric series and which has been implemented into GIS.

(Refer Slide Time: 17:06)

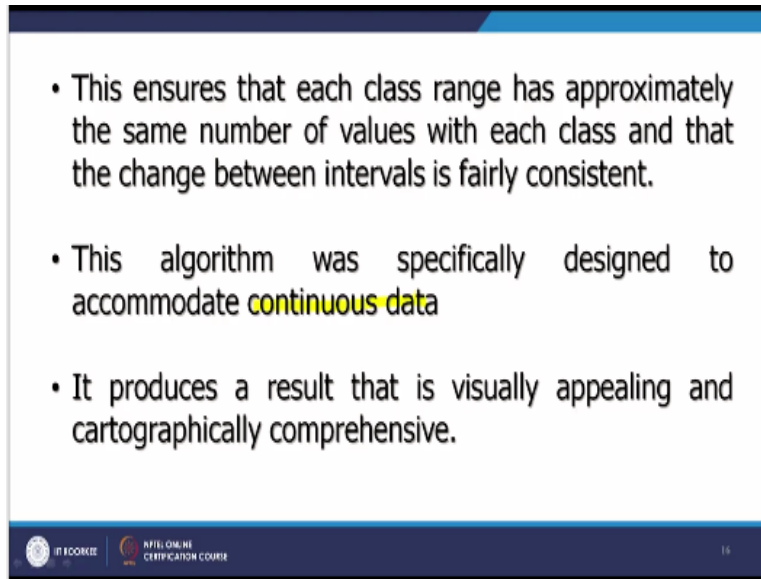
- The geometric coefficient in this classifier can change once (to its inverse) to optimize the class ranges.
- The algorithm creates these geometrical intervals by minimizing the square sum of element per class.

IT FORGE RMIT ONLINE CERTIFICATION COURSE 15

And now basically geometric coefficient in this classifier can change once (to its inverse) to optimise class ranges because you cannot or one should not go for large size of class ranges so that they group together. And this geometric classification algorithm creates these geometrical intervals again by minimising the square sum of elements per class.

So, like in Jenks; it minimises the sum of variance. Here, the square sum of variance basically. So, by which we can achieve a different classification and sometimes maybe better classification. Now, this geometrical classification also ensure that each class range has approximately the same number of values.

(Refer Slide Time: 18:10)

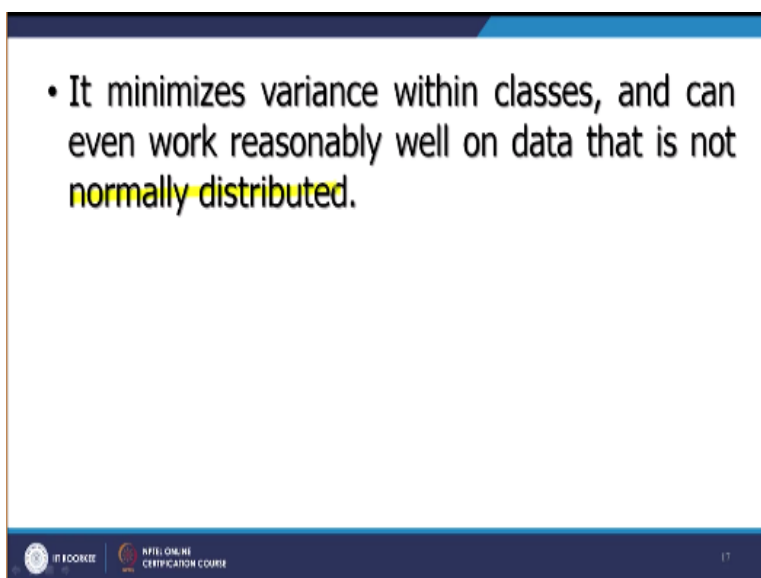


- This ensures that each class range has approximately the same number of values with each class and that the change between intervals is fairly consistent.
- This algorithm was specifically designed to accommodate **continuous data**
- It produces a result that is visually appealing and cartographically comprehensive.

So, it is basically balancing the frequency of each value or range of values in each class and that has changed between intervals is fairly constant between different classes. And this algorithm was specially designed to accommodate continuous data. So, like a digital elevation model or a satellite image which is a continuous data, this algorithm may work very well.

Or this approach of classification may work very well in many cases but again, different classification techniques may be applicable not to all but to some. Now, this geometric classification produces the result that is visually appealing and cartographically also comprehensive means people can understand much easier.

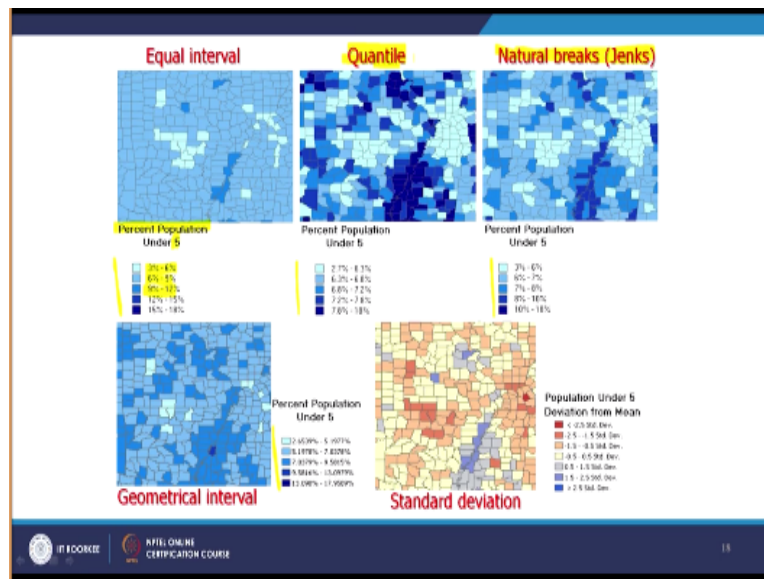
(Refer Slide Time: 18:59)



- It minimizes variance within classes, and can even work reasonably well on data that is not **normally distributed**.

Now, as mentioned that it minimises the variance within the class and can even work reasonably well on data that is not normally distributed. And you know only in ideal cases, you would find the data is normally distributed, otherwise there might be some skewness in the data and therefore, such kind of classification can give us better results. And this classification method is also called smart quantiles because it is variant from the quantile method which we have discussed earlier.

(Refer Slide Time: 19:38)



Now here the example is not from continuous data but from a vector data or a polygon data. So, what do you see that the same data set has been classified using different classification. And very interestingly out of these 5 scenarios or 5 classifications, 4 are having same number of classes that is 5. So, this is also having 5, this is also having 5, this also having 5 and this one. Except this in this standard deviation where we are having you know 7 classes.

Now if we see the equal interval, this may not give us very good results in this particular case. So, this is the population data against each land record and it is having equal interval size; that is 3 to 6, 6 to 9, 9 to 12 and so on. And of course, number of classes are 5 here. But when the same data set is subjected to a quantile classification or a natural breaks classification which is based on Jenks optimization method then it is giving completely different results.

And these classes again like in quantile; 2.7 to 6.3 and now 6.3 to 6.8 and within each class you are seeing. If you see from contrast point of view then this quantile is giving better

contrast that individual polygons or individual classes are coming very clearly as compared to equal interval. But this is data specific. It may not be true in all the cases. How the data is distributed? What is the histogram? How other statistical parameters based on that only?

So, in this particular example, quantile is giving quite good results. Then natural breaks; again 3 to 6, 6 to 7 here. So, the range is changing here. It is 3 to 6 and then 6 to 7, just 1% difference. And still, it is giving better results than your equal interval. Now, when we compare with this geometrical interval or geometrical method, again the 5 classes are there. It is again giving different results as compared to top 3 classification methods.

So, this is the point which I was emphasising that different classification methods on the same data set may give different results. And when you change the data set; apply again the same classification, you would end up with different results. So, basically the results are depending on the distribution of the data and statistics of the data. So finally, this standard deviation where you are having you know 7 ranges and the deviation from the mean here.

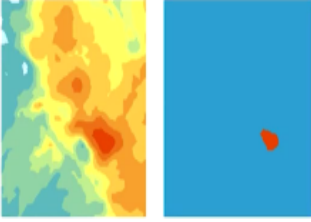
So, you are having say this one is almost mean. Then you are having you know minus values and plus values. And likewise, you know remaining 4 are having just shades of blue. But here different colours are given but colours are not important. The values are important or distribution is important. Colour; we can assign any colour. So, this also we say in GIS that colours do not matter. What matters is the values of either polygon or attribute or like cell value or pixel value, that matters.

We can assign any colour scheme to that one. So, now there is another method which is the 7th one and which is of course implemented in all GIS softwares which is custom designed classification or manual classification.

(Refer Slide Time: 23:43)

Manual Classification

Manually classification is done to emphasize a particular range of values, such as those above or below a threshold value.



For example, one may want to emphasize areas below a certain elevation level that are susceptible to flooding.

IP KOOZEE INTEL CHINE CERTIFICATION COURSE 19

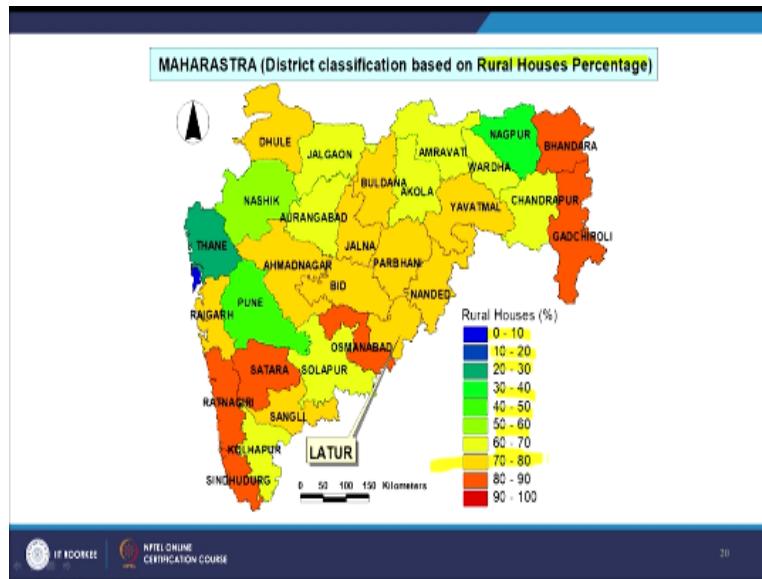
And manual classification is done to emphasise a particular range of values such as that above or below a threshold value. It is something like a density slicing as per user requirements. I want to mask everything except area which is having elevation, let us say at a particular height. So, for a hilly terrain like Himalaya; I want to see only those areas where snow falls.

So, if I know that roughly around 2000 metres and above, the snow is there. So, I want to mask everything which is below 2000 and will get only 2000 and above; something like this. The same input data is digital elevation model. It has been subjected to this manual classification or density slicing; everything has been masked except only this one. Such classifications are very useful in case of studies of flooding or other cases where we want to know which are the areas which will not get in undated, even in any kind of flooding.

And these areas then can be chosen to move people before the flooding. And this happens in the planning also. You know these natural disasters agencies are doing nowadays; they are planning the areas which are higher grounds relatively and then they will put the people there during the flooding or any other events like in a cyclone and other things.

So, through this manual classification, our ultimate aim to emphasise the areas below a certain elevation levels that are susceptible to flooding. That is the example. Or emphasise the areas which are above certain elevation for example, snow areas or the snow-covered areas. Let me give you some examples from my own work or experience, very quickly I will cover.

(Refer Slide Time: 25:50)



This is related with classification of polygons. And what you are seeing here that equal interval has been followed here; very simple, very standard one. And this classification was done on the rural house's percentage. And assuming that houses in rural areas in India are not very you know designed house or having used cement or RCC houses or columns and beams. They are just maybe made from mud and other things or a stone and other thing.

So, they are vulnerable during earthquake events. And as you know that there was an earthquake in Latur about 3 decades back and at that time you know, we analyse this data of course, after some time that and what we found that see here that 70 to 80% of the houses in Latur are rural type of houses or we say in Hindi; kutchha houses. Now, these houses are not you know engineered houses.

And large number of people died because of houses. As you know that earthquakes do not kill. This is the houses or buildings which killed the humans, not earthquakes. What I am trying to say if I am standing in an open ground where in the surrounding, no buildings are there then even a 9-magnitude earthquake comes, I will not have any harm whatsoever.

Only for a few seconds during that event, I maybe you know having some movement or something uneasy feeling but only during that period. But if I am inside a kutchha house; earthquake comes even a 6 magnitude can damage that house and then can be life threatening also. So, this is what exactly happened that in case of Latur earthquake, 70 to 80% of houses were rural houses and therefore, lot of deaths have occurred.

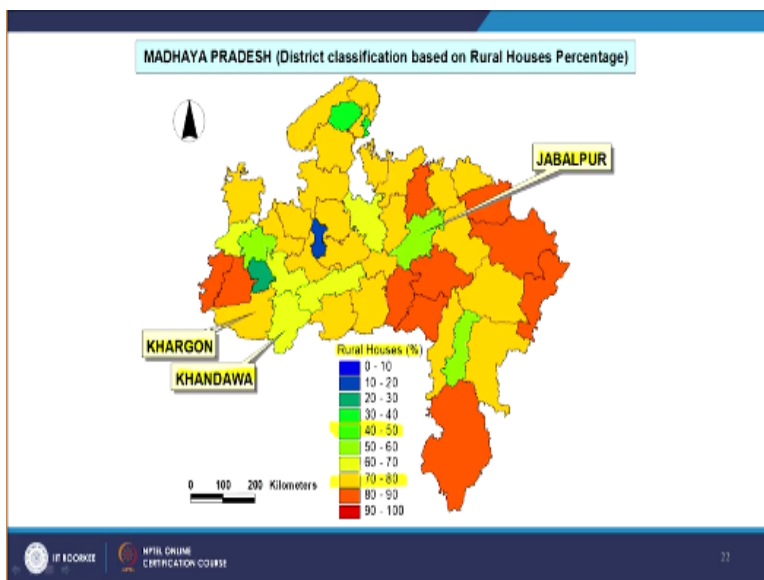
(Refer Slide Time: 28:02)

- *...there has been lot of discussion about Latur Earthquake....*
- **Why there is not much discussion about Jabalpur?**

0

IP COORISE MPIL ONLINE CERTIFICATION COURSE 21

(Refer Slide Time: 28:04)



Now, when we compared this data about the Jabalpur. Later on in Madhya Pradesh, it has occurred in Jabalpur and you see the percentage houses. Here, what we are getting the rural houses percentage is really low around 40%. Instead of in case of Latur, 70 to 80% were rural houses, here much lower in a percentage of rural houses.

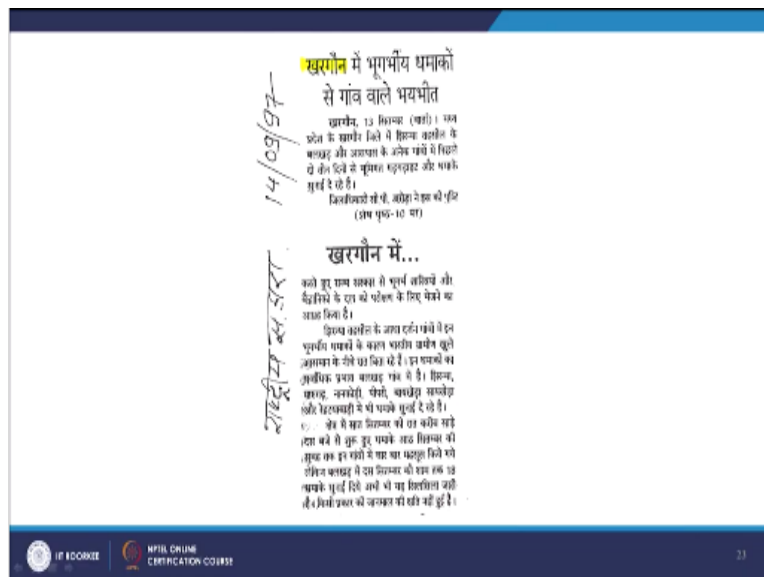
Almost same magnitude earthquake also occurred in Jabalpur after few years of this Latur earthquake. And there was hardly any death because of the number of houses or percentage of houses were urban houses rather than rural houses or where engineered houses. So even a simple data set through such classifications, we can highlight these points. Now further, if we study this, what do we find that because there is a fault which is Narmada zone fault lineament which was active during Jabalpur earthquake also.

So, after that earthquake, people talk that if further an earthquake occurs in a progression or propagates along that fault line then Khandawa and Khargon; these 2 districts are further vulnerable. If there were an earthquake, you can think that what would happen that 70 to 80% of houses are rural houses means kutcha houses. Luckily, that earthquake has not occurred.

So, what I am trying to say that using simple classification techniques even an equal interval on the real data sets; this is based on the census data. On real data sets, we can identify the problems or the causes; why it has happened one and how to plan? So, suppose the earthquake; earlier it has occurred in Jabalpur. Now in future, suppose it occurs on Khandawa or Khargon. Now this is the time to prepare people.

Because we know currently that they are having 70 to 80% rural houses or kutcha houses. In case of Jabalpur; hardly any person died because collapse of house. They were hardly any collapse. But if it happens in Khargon or Khandawa, large number of populations will die because of collapse of house or damage to the houses as happened in case of Latur. So, simple classification methods can really help us to find out not only the causes but also, it can help in the planning stages.

(Refer Slide Time: 30:59)



And this is what I was mentioning that after the Jabalpur earthquake, lot of information came. This Jabalpur earthquake occurred in 1997. I was you know forgetting the year. Later on of course, luckily the earthquake did not occur but this is how.

(Refer Slide Time: 31:14)

DISTRICT	LATUR	KHARGON	JABALPUR
THEME			
RURAL POPULATION (%)	79.6	85.0	54.5
URBAN POPULATION (%)	20.4	15.0	45.5
RURAL HOUSES (%)	79.6	83.9	55.6
URBAN HOUSES (%)	20.4	16.1	44.4

So, this is just everything in summary that rural population is this much, urban population is this much in different districts. And if you compare with Jabalpur, they are in the worst situation especially the Khargon where the seismologists thought and rural population or rural houses about 84%, just compare 56% roughly of Jabalpur.

And 79% or 80% in Latur and 1000s of people died in Latur. And if the similar magnitude earthquake would have occurred in Khargon, a greater number of people have died compared to Latur because of just poor construction of houses using like mud or stones. So, a simple classification can give us really insight about the data and application and can also explain the causes and can also help in planning. So, with this, I end this discussion. Thank you very much.