

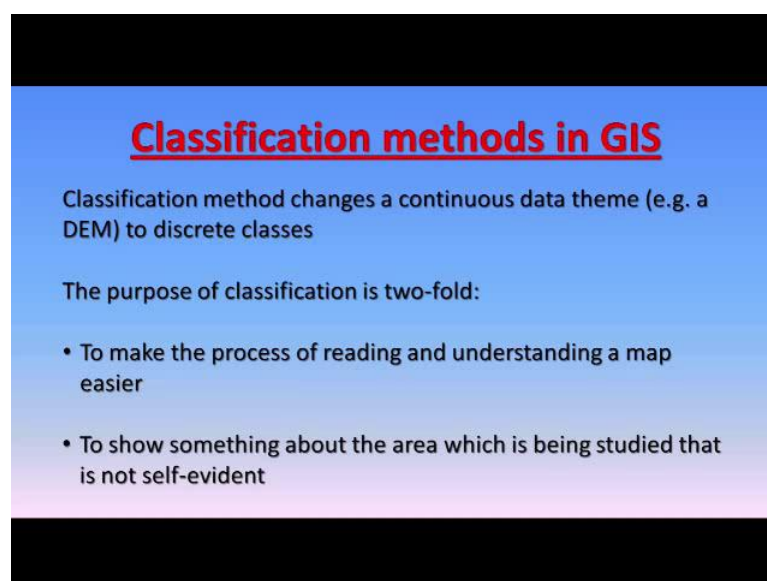
Introduction to Geographic Information Systems
Dr. Arun K Saraf
Department of Earth Sciences
Indian Institute of Technology, Roorkee

Lecture - 18
Classification Methods

Hello everyone. Welcome to the new lecture and in this lecture we are going to discuss Classification Methods in GIS. And why basically in classification, because there is sometimes you might be having continuous data and you want to discretize the data. That means, you reduce number of classes for example you are having you might be having digital elevation model where digital elevation values might be varying between say 1000 meter to 5000 meter.

Now you want to give certain categories or you want to create discreet classes for this map, say relief map kind of thing then you go for discretization that is classification. There are several methods which are available, 6 methods so far have been implemented in extended GIS softwares so we will discuss all those 6 methods very quickly. Basically the purpose here is changing a continuous data into a discreet data. That does not mean that it can only go with the raster data, it can also go with the vector data especially the polygon data and examples we will see from both raster as well as polygon data.

(Refer Slide Time: 02:03)



Classification methods in GIS

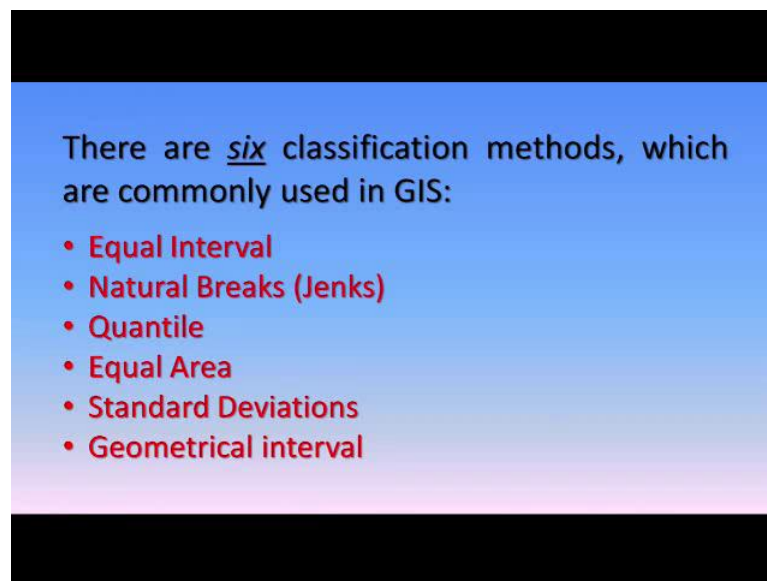
Classification method changes a continuous data theme (e.g. a DEM) to discrete classes

The purpose of classification is two-fold:

- To make the process of reading and understanding a map easier
- To show something about the area which is being studied that is not self-evident

Now the basically as I have mentioned that the purpose of classification in GIS is first is to make the continuous data in to discrete data, so that it becomes easier for reading and understanding a map and to show something about the area which is being studied that is not self evident. Sometimes you would like to hide certain things and highlight certain things and you want deemphasize something and for that purpose also the classification can be used.

(Refer Slide Time: 02:41)

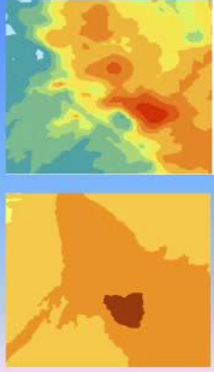


There are 6 classification methods and one by one we will go that; equal interval, natural breaks, then quantile, equal area, standard deviations, and more recent it has been added is geometrical interval which is quite useful.

(Refer Slide Time: 02:58)

Equal Interval

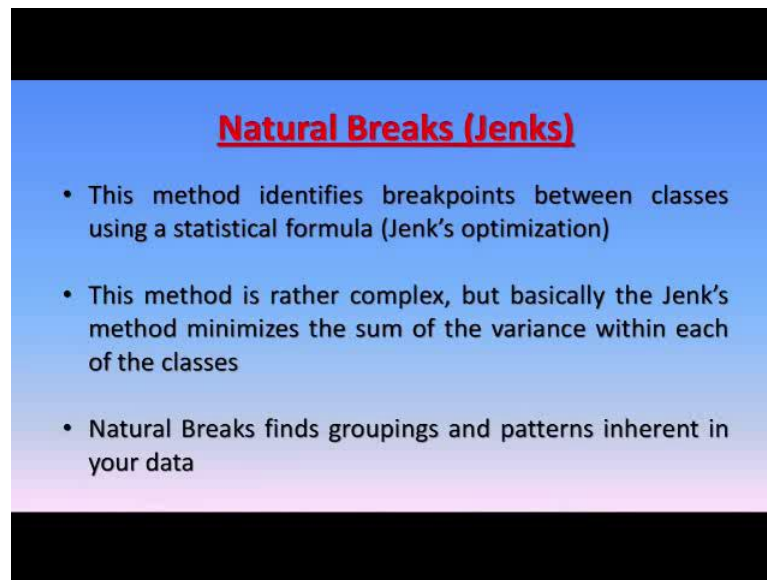
- The equal interval method divides the range of attribute values into equal sized sub-ranges
- Then the features are classified based on those sub-ranges



So, let us have first equal interval. Example is shown here there is a data which is in continuous fashion a digital elevation example is here and at equal interval it has been divided. Then it has got now the three say classes in this particular example or rather four classes. And the appearance of the continuous data into four class then it is discretized it changes completely, and therefore the interpretation has to be accordingly.

As name implies that equal interval, that means if you are having a loaf of bread and you want to slice down that bread, so general practice is the each slice is having equal thickness. Similarly here that each class I having equal range of pixel a range of your values in this particular example the cell values. So, the equal interval method divides the range of attribute values in the equal size of ranges then the features are classified based on those sub ranges. And this is how the equal area is.

(Refer Slide time: 04:19)



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

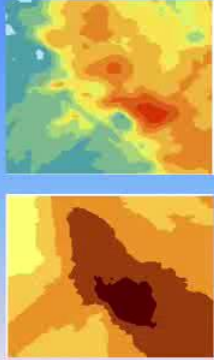
The next one is the natural breaks or it is based on the Jenks formula or Jenks optimization technique. This method identifies break points. As name implies natural breaks, it identified natural breaks between the classes using a statistical formula which is based on Jenks optimization technique. This method is not as simple as equal interval this is little complex, but a basically it uses this Jenks method of minimizing some of variance within each of the classes. And a natural break finds groupings and pattern inherent in your data.

And there this natural breaks concept is not only in GIS, but also exist in case of grading of marks of a student's. Exactly in the same way it is generally it is done that a natural grouping are find and using the natural breaks and then a for different groups and different grades are assigned.

(Refer Slide time: 05:27)

Quantile

- In this method, each class contains the same number of features
- Quantile classes are perhaps the easiest to understand, but they can be misleading.



The third one in this category of class classification from continuous to discrete is the quantile method. As that method, they each class contains the same number of features basically say based on frequency. So, the quantile classes are perhaps the easiest to understand, but they can be misleading. Example here is that the same input a digital elevation model is easier and then few classes are been generated here. So, now it may give a completely different appearance of a continuous data. So, that is why it is mentioned that it may be misleading.

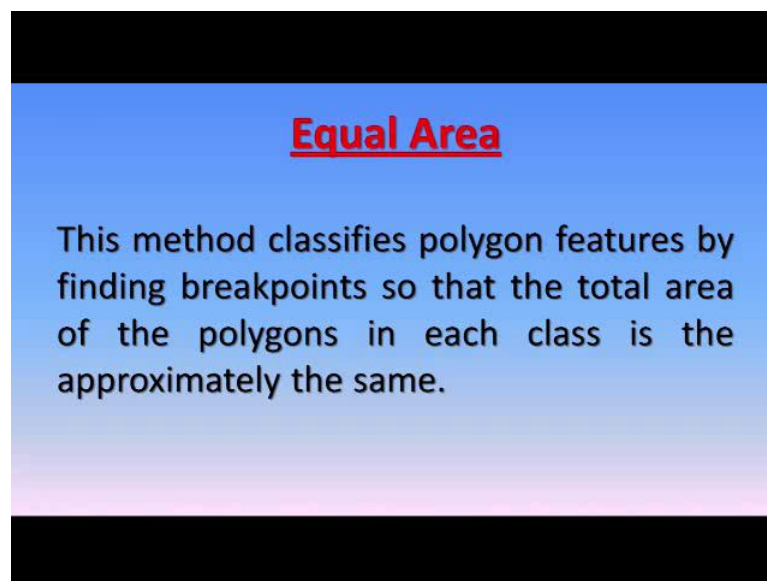
(Refer Slide time: 06:03)

- 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

Now, like in if we taken in as a for polygon example the population counts as opposed to the density of 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. If you increase the number of classes probably you are going backward. So, for all classification techniques are not suitable for all types of data it depends on the data, and therefore appropriate classification technique has to be adopted. If it is a digital elevation model then it depends on your requirement one and what kind of variations of elevation values exist with any file with in any d m accordingly you will choose the appropriate classification technique.

Quantiles are also best suited for data that is a linearly distributed and generally may not be the data may be linearly distributed may be in a Gaussian fashion or a normal distribution then accordingly a different classification technique has to be adopted.

(Refer Slide time: 07:14)



Now the fourth one is equal area. This method classifies polygon features by finding breakpoints so that the total area of the polygon in each class is appropriately the same. As a name implies equal area, so instead of having equal interval now you are looking the equal area and they in polygon it is possible to achieve this thing.

(Refer Slide time: 07:38)

- Classes determined with the equal area method are typically very similar to Quantile classes when the sizes of all the features are roughly the same.
- Equal Area will differ from Quantile if the features are of vastly different areas.

These classes are determined with the equal area method are typically very similar to quantile classes when sizes of all the features are roughly the same. And equal area will differ quantile if the features are of vastly different areas.

(Refer Slide time: 08:01)

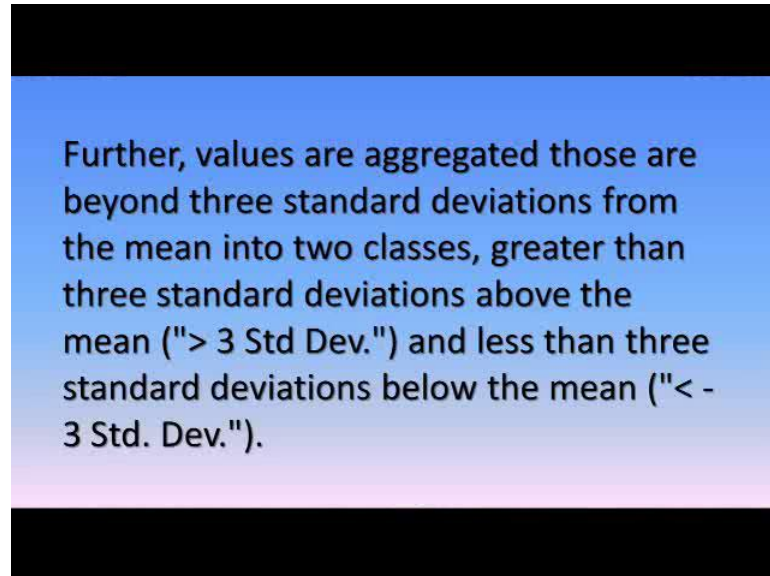
Standard Deviations

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 deviations are placed until all the data values are contained within the classes.

Now this fifth one is a standard deviation. A standard deviation in this method the mean value is formed and then class breaks above and below the mean at the interval of either one forth, half or one standard deviation are placed until all the data values are contained

within the classes. And this is a very standard practice in case of a standard understanding in case of standard deviation, so that is also employed.

(Refer Slide time: 08:28)



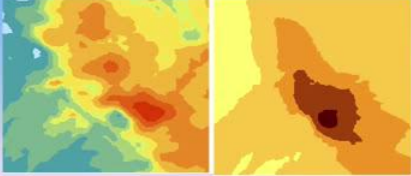
Further, values are aggregated those are beyond three standard deviation from the mean into two classes, greater than three standard deviation above the mean and less the three standard deviation below the mean. So, likewise you get.

We will see almost all examples together and see the differences how they appear differently, the same input map how they differ while using different classification technique.

(Refer Slide time: 08:42)

Geometrical interval

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



Last one in this a series of classification technique that is the 6th one is the geometrical interval which has been recently implemented in the GRS softwares and this classification a scheme where the class breaks are based on class intervals and that have a geometrical series. Example is here which again the same map is when it is subjected to geometrical interval may have a different appearance.

(Refer Slide time: 09:34)

- 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.

Now here in this a geometrical classification, the geometric coefficient is the classifier can change once to it is inverse to optimize the class range, and algorithm which creates

these geometrical intervals by minimizing the squared sum of elements per class. That means the variation within classes would not be much that is the main aim here.

(Refer Slide time: 10:04)

- 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.

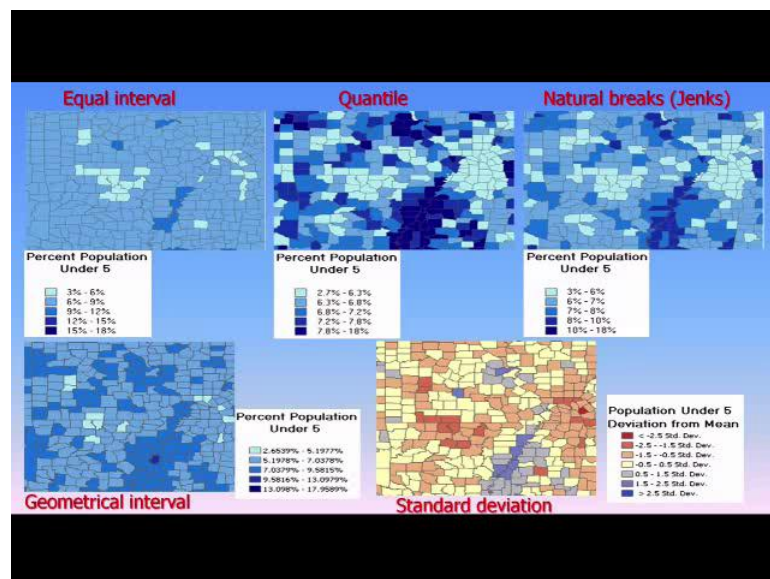
And this ensures that each class range has appropriately the same number of values with each class that the changed between interval is fairly consistent. This algorithm was specifically designed to accommodate continuous area; especially for continuous area this is very suitable. It produces a result that is visually appealing and cartographically comprehensive. That means, that not only the output maps through this analysis looks, but cartographically also they are fine.

(Refer Slide time: 10:42)

- It minimizes variance within classes, and can even work reasonably well on data that is not normally distributed.
- This classification method is also called Smart Quantiles

Now, it minimizes variance within the class as already mention and can even work reasonably well on the data that is not normally distributed. So, when data is normally distributed then again one should choose another type of classification technique and this classification method is also called a Smart Quantiles. It is a variant of basically quantile classification technique.

(Refer Slide time: 11:10)



Now the five examples are here together; one is equal interval, equal area is not here equal interval, quantile, natural breaks which is based on Jenk's optimization,

geometrical and standard deviation. As you can see here that a see this when equal interval has been chosen the number of classes in all four examples except in a standard deviation and there are a number of classes are seven otherwise here all number of classes are five. If we compare these four equal interval, quantile, natural breaks and geometrical seed by appearance of the same polygon map is completely different though number classes are constant they are five only, as you see that in equal interval that 3 to 6 one class 6 to 9 is another 9 to 12 and so on so forth.

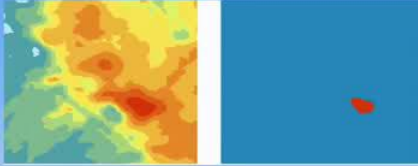
But here in quantile this is changing depending on the basically the frequency. In natural breaks again groupings have been identified and accordingly these classes have been assigned. In geometrical interval as discussed the technique has been used and a standard deviation so you are having a standard deviation here which is all at the center point is 0 and then a minus plus values are here and different class. So, input map is same, input polygon map is same different techniques will produce different results though number of classes might be same. This one has to remember.

The original one important thing here to mention is that original data will remain intact this is only for your visuals, original data will not change. Of course, once you are satisfied with the certain type of classification you can create a separate data set or you can save that particular classification technique associated with that data set. So, whenever next time you want to use you just open that small file which is having this classification information rather than creating a new data set with new classification based on new classification. Apart from these 6 classifications manual classification can always be there, that instead of going for computer to decide the classes you can also decide and which also called manual or custom classification.

(Refer Slide time: 13:56)

Manual Classification

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



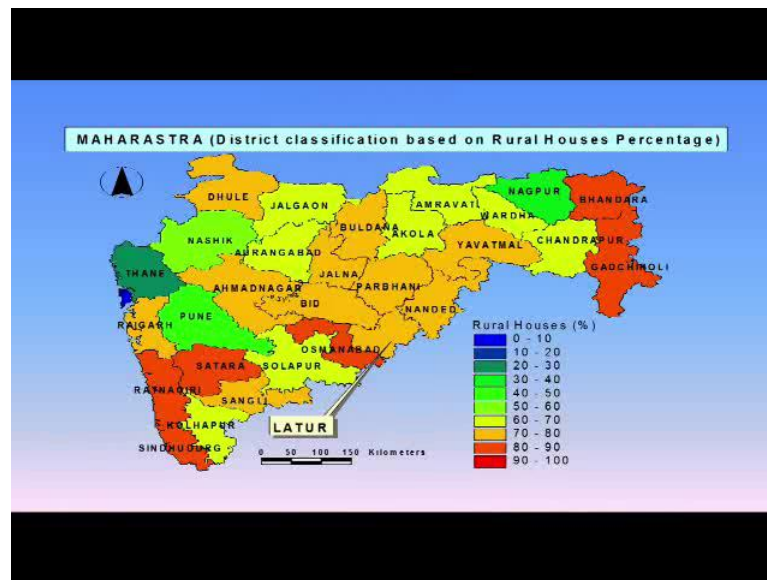
For example, one may want to emphasize areas below a certain elevation level that are susceptible to flooding.

That a manual classification is done to emphasize a particular range of values such as those above or below threshold value. Like for example, in this input digital elevation model at very high elevation points where required to be highlighted and rest were and a not required at all. May be because of in flooding if condition occurs the know that water can reach to a up to certain level and so what are the areas which would remain safe during that kind of flooding is reflected here and rest of the area will be get inundated.

So, that is why manual classification can also be employed to highlight a specific areas a specific elevations range and then de emphasize the others which are not ready useful in that sense. For example, here that one may want to emphasize areas below a certain elevation level that are susceptible to flooding. Now I will give you one example of classification techniques in the real example how it can be used.

And this we have use some years back using. Sensors data and that data and related way then earthquake event and how we can look towards the future, how this data can be used having a not information about earthquake and how we can look towards the feature. This is how one can employ and the classification techniques as well as data and some futuristic scenarios can be presented.

(Refer Slide time: 15:45)

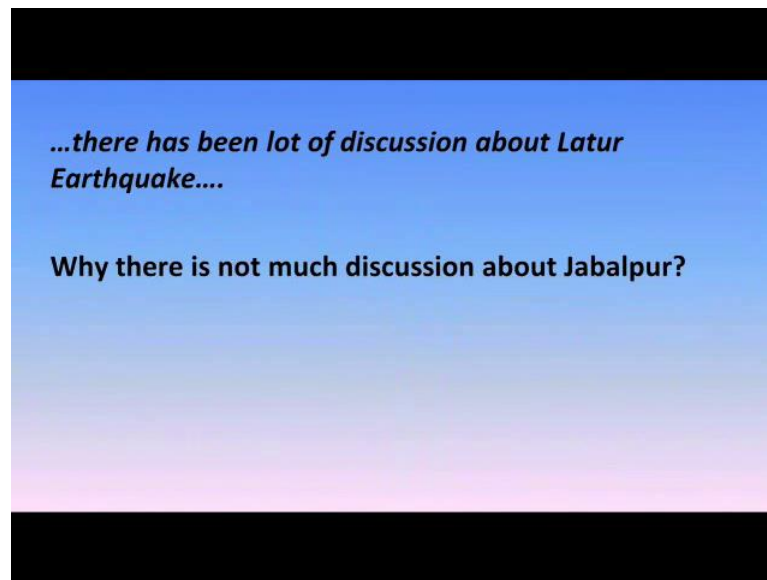


This example is based on an earthquake which occurred in Latur. As you can see that these are the rural houses in percentage and this is equal interval classification. What basically are rural houses, means there are two types of houses when the census is done and this is recorded whether it is an engineer house which is having say columns and beam and steel and other things or made from mud or bricks are called as stones. We put those houses as rural houses and engineer houses are one house.

And you see that in Latur that category even in equal interval classification it comes under this 70 to 80 percent. So, say 75 percent of the house is in this a district where rural houses and lot of large number of deaths have occurred during that Latur earthquake of 1993.

Now, later on what it was found that when this earthquake occurred people some seismologist predicted because this occurred along this Narmada soon lineament, so people predicted that there might be some propagation and then earthquakes in some other districts may also occur.

(Refer Slide time: 17:21)

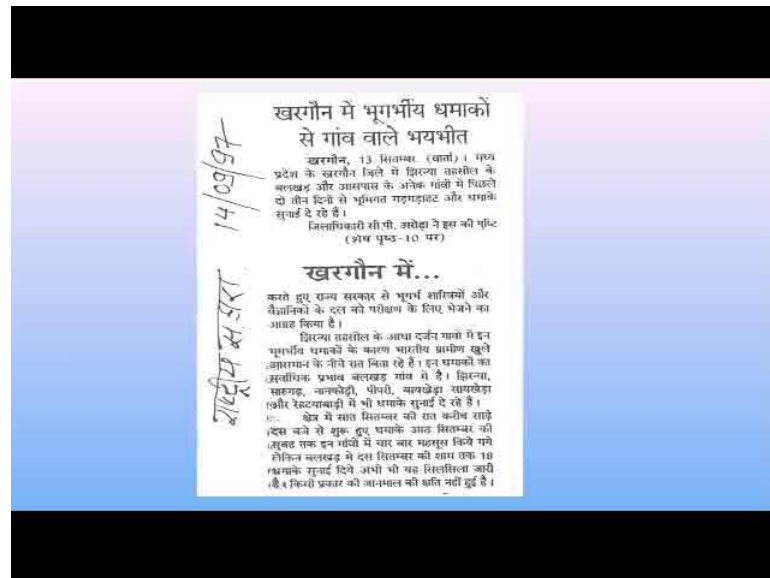


So, that that was the main discussion. Later on along on the same lineament, another earthquake occurred which is in Jabalpur, and see the same way the data was analyzed. Here what you see that is a percentage of rural houses are relatively much less, it is about 45 percent. Because of majority of houses where engineered houses and therefore there were only few deaths. Relatively in case of Latur the number of houses in a rural category where of about 75 percent, and therefore large number of deaths.

So, using a simple data set one can analyze data what has happened in case of Latur and what would happen in case of in some other district if at all earthquake occurs. For example, that earthquake occurred in Jabalpur, if that would have occurred in other districts neighboring this districts then again we might have reach to the same roughly 65 to 75 percent. And this is what the predicted that in this Narmada so lineament is going and the earthquake might occur in these districts. If those would have occurred say in Khargon then 70 to 80 percent of the house where rural houses number of death would have been almost same as in case of Latur.

Even if it is Jabalpur earthquake instead of occurring here along this Narmada son lineament if it would not have occurred on in Khargon district then similar scenarios like Latur would have happen, but luckily instead it occurred in Jabalpur where number of rural houses are relatively less compare to engineer house.

(Refer Slide time: 19:19)



So, using a simpler concept you can think that this is what the prediction were there, that now the next earthquake after Jabalpur would occur in Khargon and this Jabalpur earthquake occurred in 1997.

(Refer Slide time: 19:35)

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 the scenario that in rural population, urban population, rural houses and urban houses and rural houses and Latur where about roughly of 80 percent, whereas in case of Jabalpur just 56 percent, and few deaths, thousands of death. If the similar magnitude

earthquake instead of Jabalpur or instead of Latur would have occurred in Khargon then 84 percent of houses were at that time a rural house.

So, relatively if assuming that same density of populations, same number of people exists in that area as in Latur more number of deaths would have occurred. Now using a simple analysis in GIS and the data which almost available freely through the census department such scenarios can be presented to decision makers why, because this is what the future as to we do because now we know that it is all this Jabalpur Khargon all these districts are falling all along the Narmada son lineament.

Recent earthquake in 97 Jabalpur has occurred, in future there might be some earthquakes may occur all a in any of these districts if at all it occurs then some majors and probably there is a time to take certain remedial measures, educate the people and may go for a (Refer Time: 31:06). So, it is starting from a simple classification technique even equal interval can be implied to look these scenarios which can give us some insight of the what is going to happen in future.

And that brings the end of this particular presentation on the classification techniques in GIS for continuous data to convert them to discrete data.

Thank you very much.