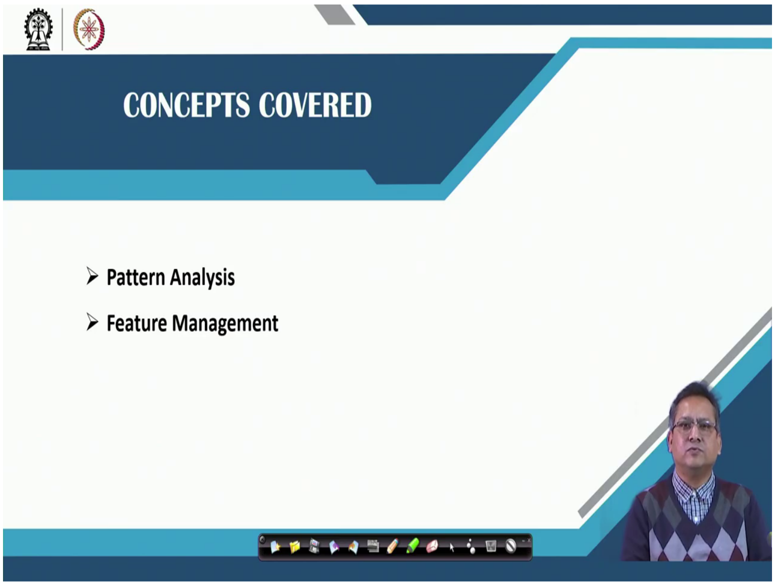
Geo Spatial Analysis in Urban Planning
Prof. Saikat Kumar Paul
Department of Architecture and Regional Planning
Indian Institute of Technology, Kharagpur

Lecture – 08
Vector Data Analysis (Contd.)

So, welcome back dear students; to the lecture 8 which is a continuation of the previous lecture of Vector Data Analysis.

(Refer Slide Time: 00:21)

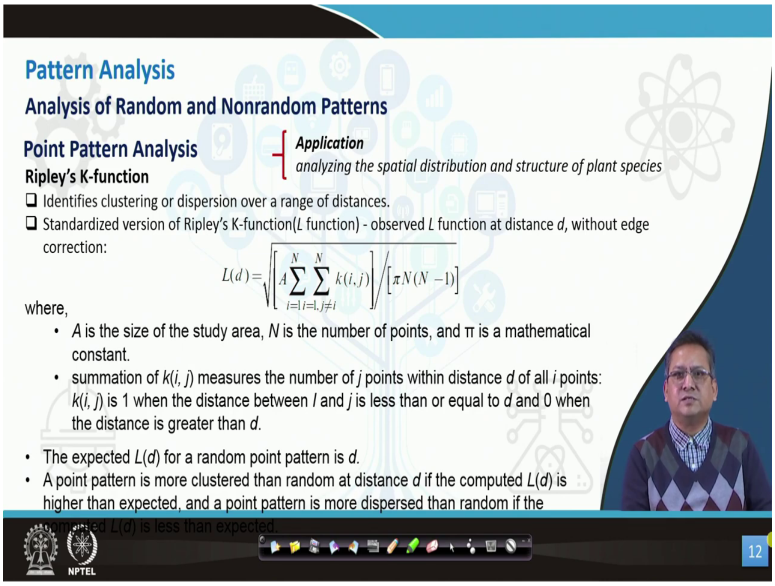


So, in the earlier lecture we had ended with the pattern analysis where we had talked about the I mean the neighborhood pattern to identify the pattern.

So, today we are going to see, I mean in this particular lectures we are going to see the different ways of doing the pattern analysis other ways of doing the pattern analysis; what happens when the values are different, whether we can identify the patterns, whether we can, how we can do the patterning at a global level or at a local level.

And we are also going to see the different operations related to feature management and how they can be used in your ah concept of urban planning whenever you are doing some applications how you can use that.

(Refer Slide Time: 01:07)

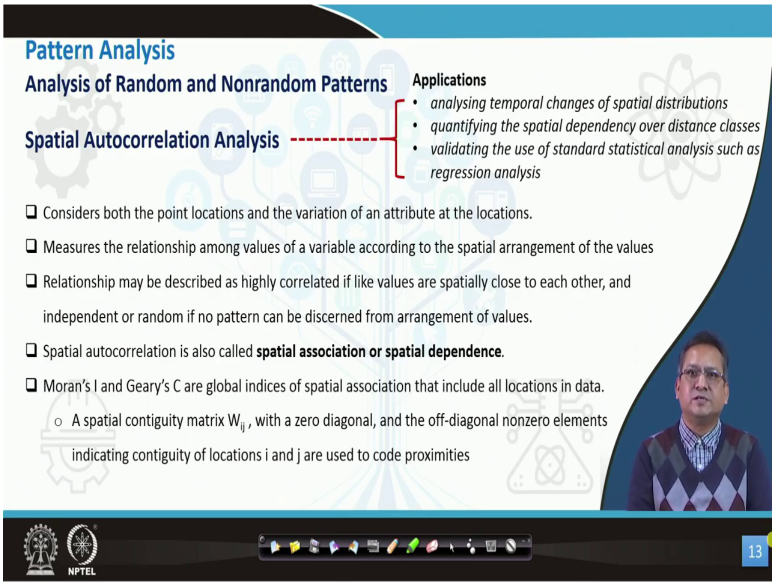


So, in the pattern analysis let us look into the analysis of random versus non random patterns. So, we use our Ripley's K function your which would basically identify or find out the dispersion. If we have a distance as a function so over a range of distance whether I mean there is a clustering or dispersion. So, this can be identified using the Ripley's K function.

Now, the standard function of Ripley's K function is also known as L function it is the observed L function. So, which is observed at distance d, without doing a edge correction. Now, this function L is calculated using this particular equation which is shown out here where in, in this case your A is the size of the study area and N is the number of the point the capital N is the number of points and pi is the mathematical constant where the summation of k measures the number of points within the distance d of all points.

So, where k i, j is 1 and when the distance between say i and j is less than or equal to d and 0 when the distance is greater than d. So, apart from the Ripley's function, we can get the expected value of the random point pattern which is d. And the point pattern is considered to be a to be more clustered than random at distance d. If the L d score is higher than expected and the point pattern is more dispersed than random if the Ld score is I mean it is less. So, than the expected.

(Refer Slide Time: 03:16)



So, apart from that we can also work out the spatial auto correlation analysis apart from the Ripley's coefficient L d function we can also work out the spatial auto correlation between the points. So, we find out the variation of the attribute I mean at the different locations. It would measure the values basically I mean according to the spatial arrangement and it would try to correlate the like values I mean highly correlated or like values would be identified.

And also the values of these points which are close to each other, so it could be either independent or random. I mean, if no pattern is discerned I mean if you cannot see any sort of a pattern emerging out of it, then it would be independent or it could be considered as random. Now, it is also I mean referred to in some of the literature as spatial association or spatial dependence.

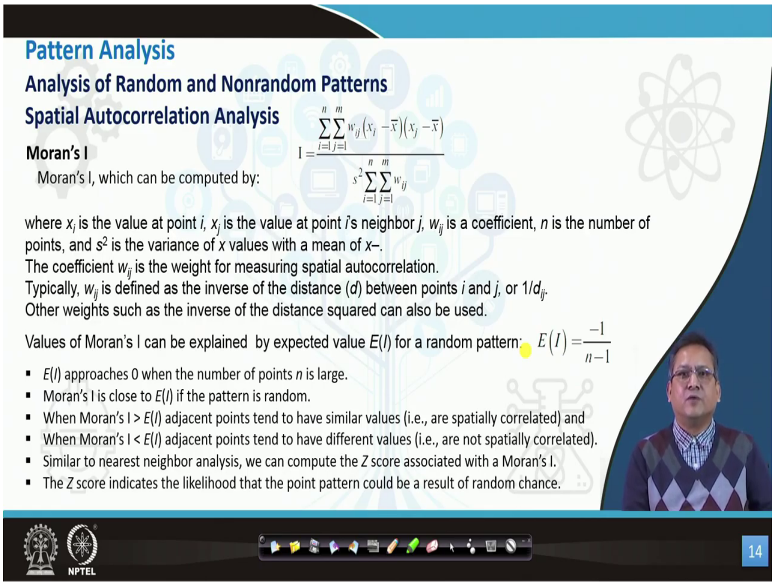
Now, there are two indices which are very extensively used in GIS which is known as Moran's I and Geary's C which are the global indices I mean these are not the local metrics, but the global indices of spatial association which include all locations in the data. So, we also have local versions of this Moran's I and Gearys C which we will look into it. Now, we work on the spatial contiguity matrix having a zero diagonal.

So, all the elements of these spatial contiguity matrix would have zero values in the diagonal elements and the off diagonal elements are nonzero values which would indicate the contiguity of locations i and j. And, basically we use this non off diagonal elements to code the proximities I mean how the points are related to each other.

Now, there could be applications I mean where can we apply this spatial auto correlation analysis we can do it for analyzing the changes over time the temporal changes of the special distributions. We can also use it to quantify the spatial dependency over a distance classes I mean if we have a distance I mean a metric in a as one I mean as one requisite.

And we want to see what is the spatial dependence of some of the points lying in proximity to the particular metric we can quantify that using spatial auto correlation analysis. We can also validate the and use the standard statistical analysis such as the regression analysis. So, when we get outputs from regression data we can also validate it cross validated using the spatial correlation analysis.

(Refer Slide Time: 06:32)



Now, we are talking about the spatial auto correlation analysis in which we had talked about the Moran's I and Geary's C. So, let us see what is Moran's I and how it can be used. So, it is basically computed using this particular equation which is a w ij x i minus x bar. And we do a sigma I mean it is a nested operation from i and j starting from 1 to values of m and n and we multiply this x i minus x mean into x j minus x mean divided by the standard deviation into a the summation of your w ij values.

Now, x ij is the value at point i x i is the value at point i. x j is the value of a given I mean theme at point j I mean point i is neighbor which is point j and w ij is a coefficient. Now, n are the number of points; the small n are the number of points and a square is the variance of x values with mean of x bar.

Now, it the coefficient w ij is used for measuring the spatial auto correlation. Now, it is the w j ij it is a that inwards distance between the point i and point j or it can also be written as 1 by d subscript ij. So, the other weights the that is the inwards distance squared or can or those kinds of weights can also be used here, when we are talking about weights. So, the values of Moran I what it can explain; it can explain the it is explained by the expected value for the random pattern which is given by minus 1 divided by n minus 1.

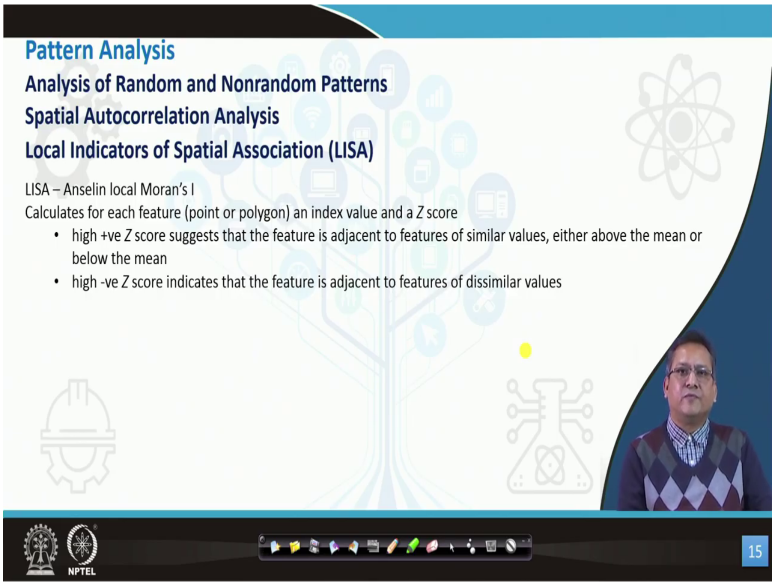
Now, n is the number of the points when I mean if the n is a very huge, then what would happen is this E I would approach 0. Now, the Moran's I is close to Ei if the pattern is random. I mean if we calculate the value of i from this earlier equation that we have seen over here. So, if we calculate the Moran's I so, this value is close to the expected EI value which is the expected value in case if the pattern is random.

Now, when the value of Moran's I is greater than this expected value then what happens is this points tend to have similar values and they are spatially correlated. And, if it is the other way around if the Moran's I is less than EI, then I mean the adjacent point tend to have different values.

So, if you have the Moran's I higher than your expected values you can safely assume that the points would be clustered and they would have values which are similar in nature. You may you would have similar sort of features a cluster together, but when the Moran's I is less than the expected values then they are I mean they would have different values in the clusters and they are also not spatially correlated.

So, it is similar to the nearest neighbor analysis that we had done wherein we had computed the Z score associated and I mean we can do the same thing we can compute the Z score with the Moran's I. And in this case the Z score would give you the likelihood that the point pattern would be the result I mean of a random chance.

(Refer Slide Time: 10:39)

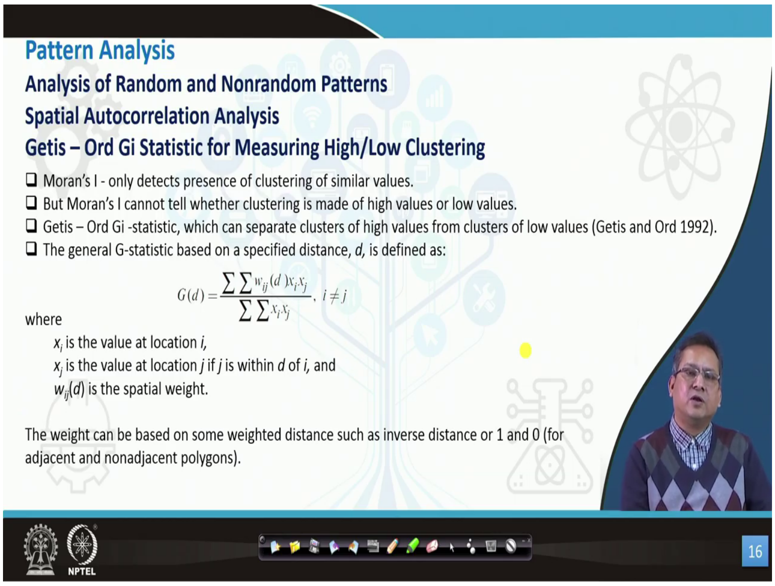


I mean you would have, I mean it could have happened randomly not in a very specifically or I mean organized manner.

So, we can also have local indicators for spatial association which is known as LISA in short, it was given by Anselin. So, it is also known as Anselin local Moran's I in literature, wherein it calculates for each point or polygon it would calculate an index value and a Z score.

Now, if we have high positive Z score, it would suggest that the adjacent features are similar in value and it is either above the beam or below the mean. Now, if the values are negative, if we have high negative values it indicates that the adjacent feature has dissimilar value.

(Refer Slide Time: 11:30)



So, next I mean since we had talked about your Moran's I let us see how the Getis Ord Gi statistic is different from Moran's I. So, it gives you the high or the low clustering we had seen that Moran's I only gives the presence of clustering of similar values. I mean we had worked out the I mean a coefficient I using that equation.

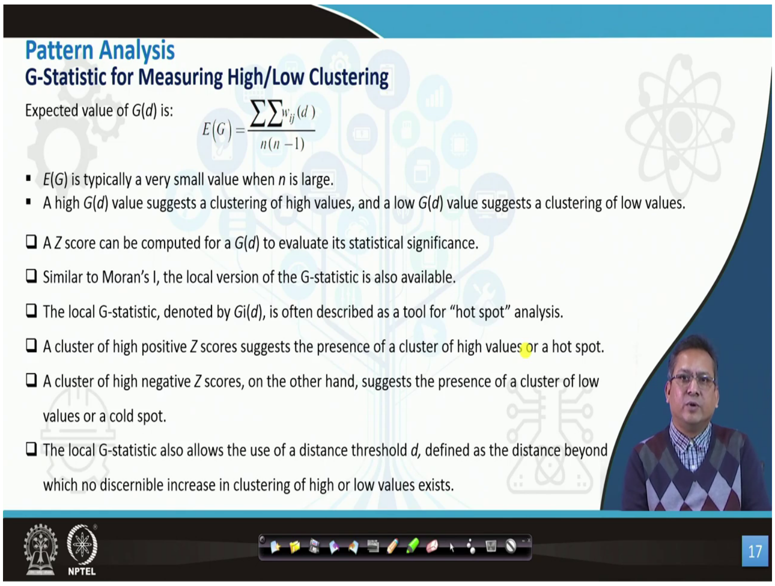
So, it only gave us the presence of a clustering of similar values ah, but it could not detect I mean a clustering I mean whether the clustering is made out of high or low values. I mean if you have differences in the values of points I mean all the points do not have same values then whether the clustering is of the high values or clustering is of the low values cannot be given by the Moran's I index.

So, in that case we use the Getis Ord Gi statistic which can separate the cluster of high values and differentiate it from the clusters of the low values. So, this was given by Getis Ord in 1992 and it is based on the specific distance d, which is given by this particular I mean equation G function of d which is summation of basically d ij into d into x ij x i into x j divided by sigma of x i into x j.

Now, we had earlier seen it is also similar to that that x i is the value of location a value at location i and x j is the value at location j with j is within d of i that is within the I mean distance d and a w ij d is the spatial weight.

So, the weight is based on some weighted distance as it could be a inverse distance or it could be a value of 1 or 0 depending on whether the polygons are adjacent to each other or whether they are not adjacent to each other. If they are not lying adjacent to each other you can code it as you may code it has 0 or if you they are adjacent to each other you can code it has 1.

(Refer Slide Time: 13:55)



Now, the expected value I mean we always talk about the expected value and we compared the your Moran's I with the expected value, similarly we can calculate the expected value for the Getis Ord Gi measurement or the G statistic measurement. So, it can be calculated using this particular equation which is a ratio of the summation of the w i divided by n into n minus 1; where n is the number of points.

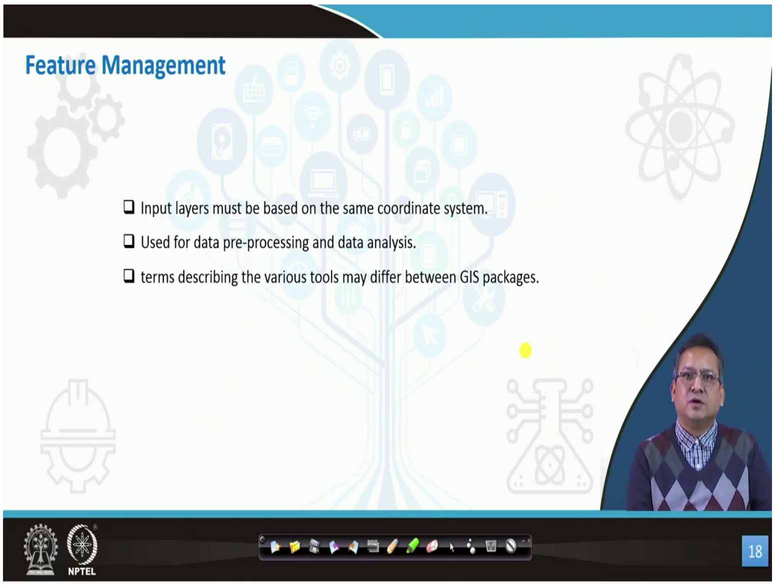
So, is EG is a your typically very small when ever your n is going to be large. If you are number of points are very large, then the value of expected clustering index would be very small. So, a high value of Gd, I mean we had talked about the Getis Ord statistic. So, if we have very high value, then it would suggest that the there is a clustering of the high values and vice versa. I mean whenever we have a low Gd value, it would suggest that the clustering is of the low values.

Now, Z score can also be computed to evaluate the statistical significance. So, it is similar to the Moran's I and it is a local version of the Gi statistic is also available. We had seen the Anselin Moran's I index. So, local version of the Gi statistics can also be measured. So, this is known as the local G statistic it is known; it is generally denoted as G i d. So, it is used as a tool for hotspot analysis.

So, a cluster of high positive Z scores would suggest presence of cluster of high values or a hotspot, so in a local domain. And a cluster of high values high negative values or Z scores on the other hand generally would predict that the cluster of low values are there or there could be a cold spot.

So, I mean the Gi statistic, basically would allow the distance threshold. I mean you can use the distance threshold d and it is defined as the distance I mean beyond which no major increase I mean no visible increase is there in terms of clustering of high or low values.

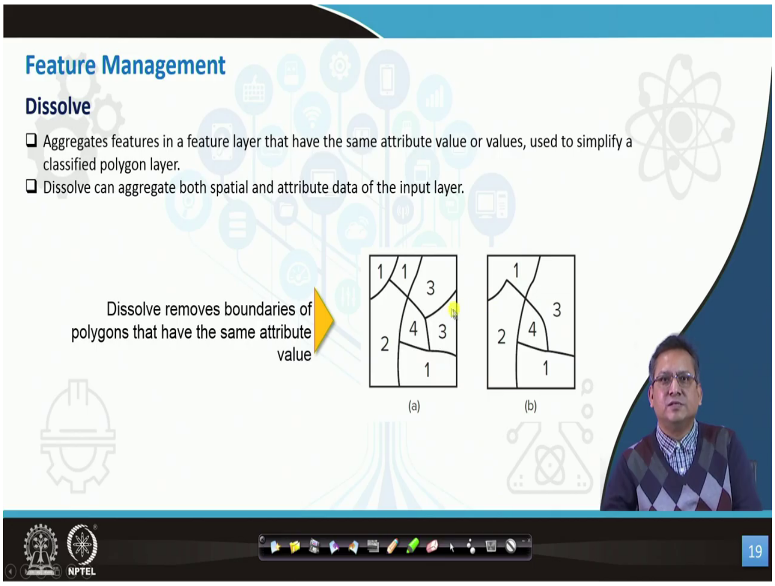
(Refer Slide Time: 16:45)



So, the local Gi statistic would give you the distance threshold where you cannot see that there is a major change in terms of the clustering of high or low values now let us get into the feature management part once we had seen the different ways to cluster, we had seen the global clustering algorithms, we had seen the local clustering algorithm.

So, let us see what are the different feature management tools that GIS has and how it can be used. So, I mean for the feature management to happen the conditionalities are that the input layers has to be on the same reference system same coordinate system.

(Refer Slide Time: 17:33)



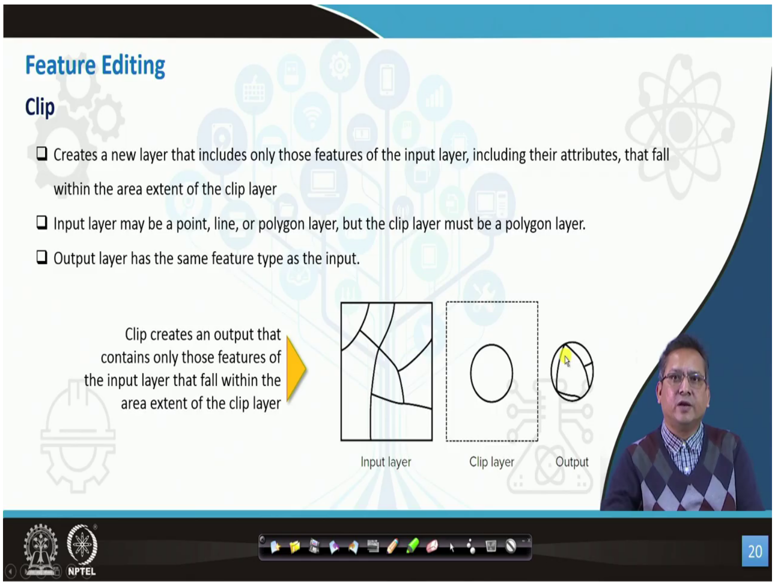
So, I mean it is used generally for pre processing or it is also used sometimes for the data analysis. So, there are terms I mean which be differ I mean these terms are not unique for the GIS tools, I mean different software's may have their own I mean terminologies or definitions or say keywords when they are talking about feature management. So, the first tool that we are talking about is the dissolve command.

Now, this command or tool basically aggregates the feature in the feature layer having similar or the same attribute values to basically, it would simplify if you have a polygon layer wherein you have multiple polygons of the same classes. So, you can do a aggregation of the different features you can aggregate it. So, I mean if the boundaries are shared then you can dissolve it into a polygon of a common boundary.

So, you can reduce the number of polygons. So, the data size will become smaller and it will make your analysis more efficient. So, I mean the dissolve generally would be able to aggregate the attribute data as well as the spatial data of the input layer. So, here as an example you can see that it removes the boundary I mean the we have two classes here; which is I mean denoted by these 2 polygons are having a class 1.

So, when we run the dissolve operation I mean both this boundary gets dissolved and you have only one attribute that attribute as well as the boundary gets dissolved. So, you have only 1 entity polygon after dissolving it and similarly for this class 3 you can see there are 2 entities which gets dissolved as 1 entity.

(Refer Slide Time: 19:26)

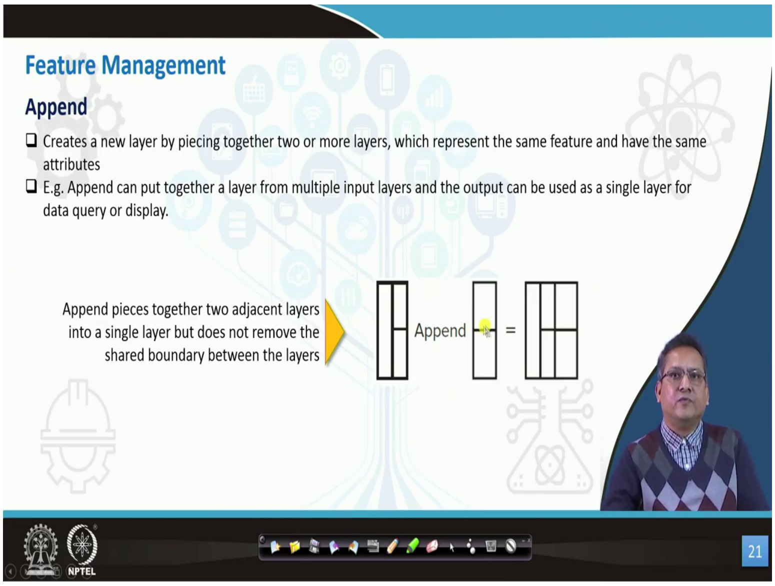


Now, the next editing feature is the clip command which is basically used to subset different layers. So, your input layer can be a line it could be a point or it could be a point a polygon layer, but the clip layer the layer by which you are going to clip your data has to be a polygon layer.

Now, the output layer would have the same feature type as the input layer. So, if your input layer is a point; if your input layer is a line that in that case your output data would be either a point or in the later part if you take it as a line, then your output data also would be that of a line. Now, you will not get those entities of the clip cover, if you have a clip layer those entities will not be reflected in the output data.

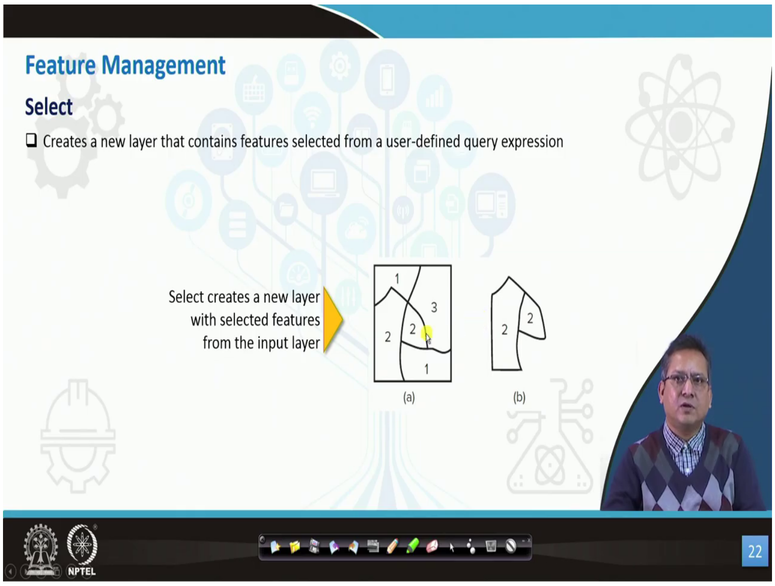
So, you can see for an example that this is your input say suppose this is an urban area and you have different land use classes in this and you may have a urban boundary which is a you can approximate it by a this particular circle for say in this particular example. So, when we clip this particular polygon cover using this clip layer you can see you only retain the entities which are there in the input layer and those entities are line entities.

(Refer Slide Time: 20:57)



So, we can also create a we have also have an management tool; feature management tool which is known as append. So, it would join two layers it would I mean create together piece together or join together two layers two or more layers and represent the same features and it would have the same attribute. So, I mean you can have multiple inputs and the output would be a single layer for data query or display. So, if you say suppose if you append this particular polygon to this I mean input, then in the output we can see that it would be joining the input layer.

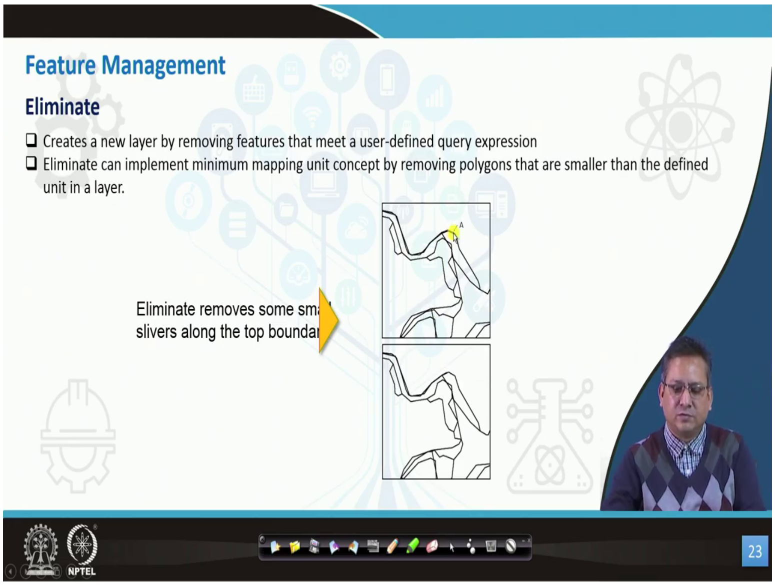
(Refer Slide Time: 21:53)



Next, we go into the select feature. The select feature basically creates a new feature layer from the selected features in a given layer. So, if you run a query if we give a query operations say suppose we have different words in a urban area having different densities and you want to pick up the words which has higher density.

So, you can run a query, so it may result in giving you word 2 as the outcome. So, in the if you take the select command it will create a new feature layer wherein you will get these only these two polygon entities that is this word 2. I mean values having 2 which might indicate a density in a given word.

(Refer Slide Time: 22:42)

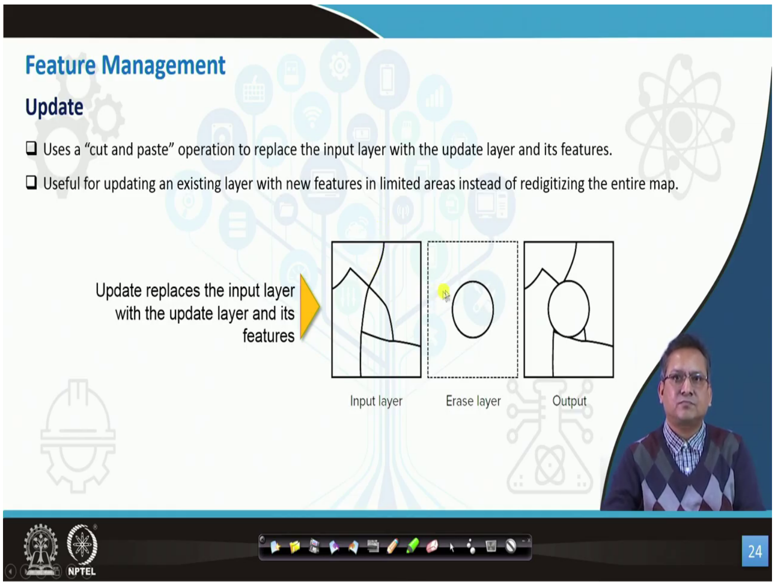


We also have a eliminate I mean tool when we are talking about the feature management; it creates a new layer by removing features that meet the user defined query expression.

Now, it uses or it can implement the minimum unit a minimum mapping unit concept by removing polygons. So, if you have slivers in data whenever say you are intersecting two polygons you may see there are resulting polygons slivers, small polygons that you may want to eliminate. So, you can eliminate by giving I mean area limitation that I mean polygons beyond this I mean below this particular area needs to be eliminated.

So, by that you can eliminate the sliver polygon. So, you can see it in this particular example that we have this small area where my cursor is there you can you see this yellow bubble. So, we have the small areas as slivers. So, when we are using the eliminate command it gets removed I mean there is a I mean small sliver at this point in it.

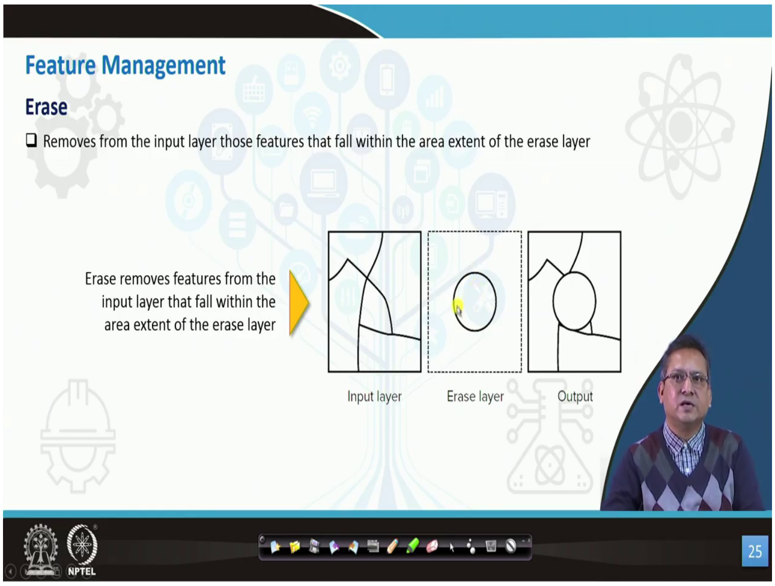
(Refer Slide Time: 24:06)



So, you can set thresholds for the eliminate command as well by setting the area of the entity. Now, we can also have do updation of the entities which basically uses a cut and paste operation, and it is it replaces the input layer with the updated layer and its features I mean it also includes the attribute layers.

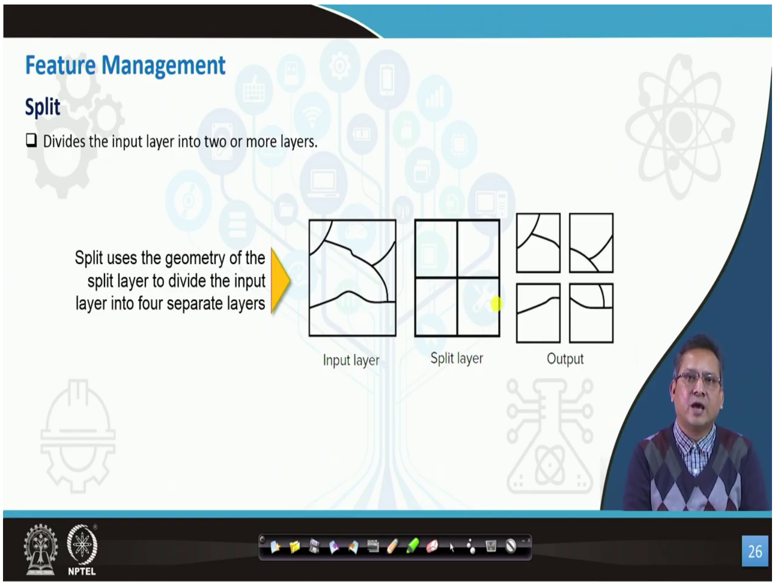
So, I mean we use this say suppose there is a change in the boundary of a given extent then we use the update command when we are editing it. And instead of re digitizing the entire map what we can do is we can update the small area where we are I mean working on that. So, you can see that this area has been updated; we have erased this particular use this erase layer. So, in this particular circle whatever information is there it is I mean erased and you see the updated one.

(Refer Slide Time: 25:08)



So, update basically works on that, then we may have the erase command which removes from the input layers the features which fall within the area extent of the erase layer. So, I mean it is also similar to the update command. So, but in this case if you have entities within this erase layer you can see that it gets erased. I mean there is no I mean information in the output image.

(Refer Slide Time: 25:37)



Split the next feature which is used for feature management it divides the input layer into two or more layers. So, if we have a polygon and we want to divide it or there is a new bifurcation of a administrative boundary and we divide it by a line.

So, I mean when we have the split layer and we superimpose on it will create the split layers. So, in this you have the split layer which is the out I mean the regulating layer and you have the input layer. So, in the output geometry I mean you can see the split and it divides the input layer into four different sets of layers in this case.

(Refer Slide Time: 26:28)

So, a recapitulation of what we had gone through in this particular lecture on your vector modeling. So, vector data modeling in GIS. So, first we had talked about the concept of buffering of points, lines and polygon.

We had talked about a singular buffer we said the buffering is a you can do it on either side of it you can do it on the right side or the left side it need not be always on the center of a line you can have the buffer the ends could be a (Refer Time: 27:07) or the ends could be circular in nature.

So, you can have different I mean ah types of buffers as well and we had also seen that the buffering distances need not to be same. I mean if you have multiple buffers you are creating trying to create multiple buffers the distance criteria could differ for the different buffering regions. And we had also seen the I mean talked about the different applications where you can use this buffering.

Next, we had seen the overlay operation wherein, we had seen how the Boolean operators are used for the overlay function. And we had seen the different methods of overlay when we were talking about the overlay function. Then we had talked about the distance measurement which is extensively used in modeling and we had talked about the Euclidean distance which is used to measure the distance from a particular feature and all the distance based models can be applied using this distance measurement function.

Then we had talked about the pattern analysis. So, in that pattern analysis we had talked about different measures, we had talked about some global measures, and then I mean we had talked about Moran's I we had talked about Getis Ord Gi and then we had talked about the local measures.

So, it is important to identify the scores I mean whether it is the Moran's I score or whether it is the Getis Ord Gi score or the Anselin index and we compare it with the expected score and the Z score to give us the idea regarding the pattern of distribution of a particular I mean a theme or particular information in a given urban area. So, these type of analysis cover most of the GIS operations and it is I mean most of the analysis that we do over vector datasets would entail using some of these operations when we are doing the modeling.

So, thanks and in the I mean next lecture we will look into the raster modeling, and we had also a in this particular lecture we had talked about the feature management. For a the feature management these are basically aided features which are there in most of the GIS package it may be named differently your overlay features and the feature management issues, but it is there in most of the GIS database which is a GIS packages which is used for the GIS analysis.

So, thanks once again.