Geographic Information Systems

Prof. BHARATH H AITHAL
Ranbir and Chitra Gupta School of Infrastructure Design and Management
Indian Institute of Technology – Kharagpur

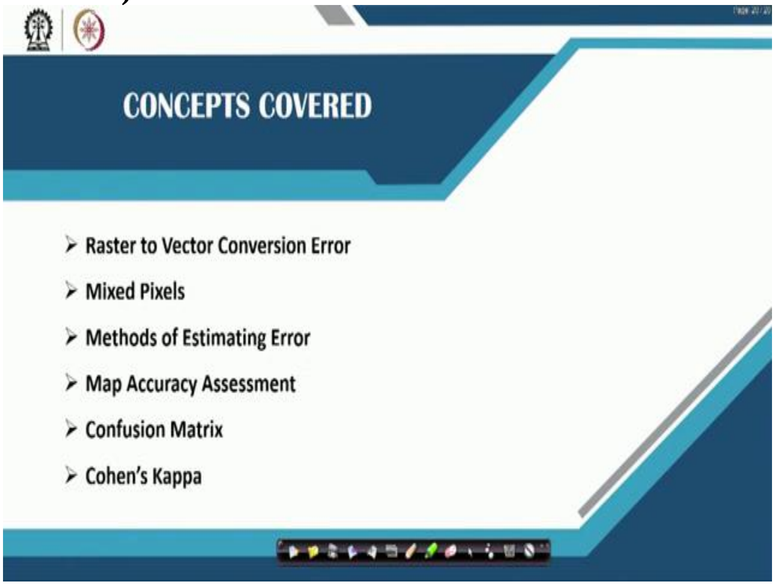
Lecture – 23
Classification accuracy and pixel errors

Hello Namaste, welcome to the next session of the module 5 where we are looking at GIS data quality. In my previous sessions I spoke about what is an error how an error can creep in your GIS data set, how you will actually try to address those errors. Then we also looked at data quality issues we also spoke that your data should be complete. We also looked at logical consistency of your data.

So, all your and attribute issues with attribute data issues with the spatial data. So, you looked at all of these aspects that are needed when you are addressing the data acquisition to the data model. Now in this part of the lecture will also extend the same discussion, but we will also look at even the raster part of it, the raster model, what are the things that may come and how do we look at the errors.

How we look at the accuracy and probably will also look at the pixel errors which are extremely important in terms of understanding and image.

(Refer Slide Time: 01:30)



CONCEPTS COVERED

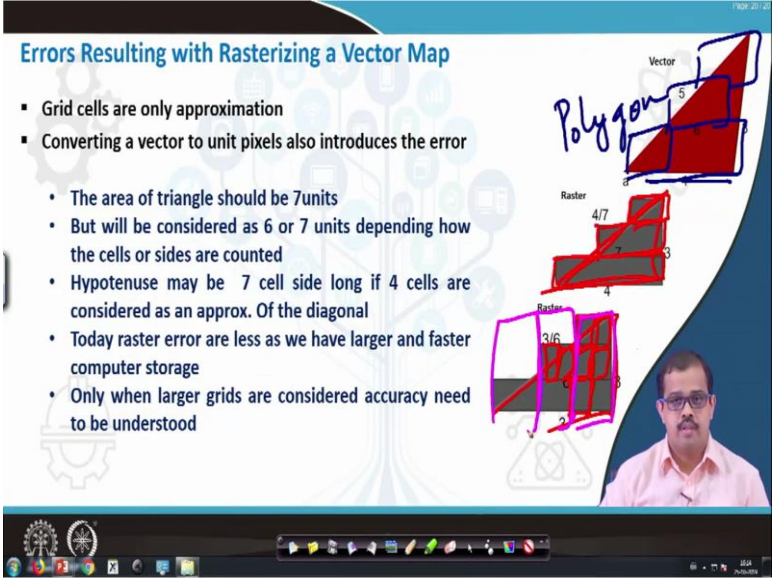
- Raster to Vector Conversion Error
- Mixed Pixels
- Methods of Estimating Error
- Map Accuracy Assessment
- Confusion Matrix
- Cohen's Kappa

So, the concepts that I would speak about today is on raster to vector conversion how do you convert a raster to a vector? So, when you are converting a raster to a vector, what are the different errors that you may occur it may come as an error then you will have mixed pixels issue for example in and let us take an pixel. So, in case there are certain issues that is there in that particular data or may be its special resolution is very low. So, you may see that a lot of different regions have clubbed together to form a single region.

So, that kind of issues may occur. So, we will address how will we look at how do we address that issues also, then methods of estimating error, what is the different kinds of error that may come using a mixed pixels, where we will look at 2 different methods, very well-known methods. Then we will look at map accuracy assessment. So when I look at that accuracy assessment, I am looking at the raster data model.

Wherein we will look at both the confusion matrix and finally ended with the kappa. So, kappa is one of the measurements of how we look at the classification system in a raster model. So, these are the things that I would cover in today's lecture.

(Refer Slide Time: 02:51)



Errors Resulting with Rasterizing a Vector Map

- Grid cells are only approximation
- Converting a vector to unit pixels also introduces the error
  - The area of triangle should be 7units
  - But will be considered as 6 or 7 units depending how the cells or sides are counted
  - Hypotenuse may be 7 cell side long if 4 cells are considered as an approx. Of the diagonal
  - Today raster error are less as we have larger and faster computer storage
  - Only when larger grids are considered accuracy need to be understood

So, when I started, the first thing that I would like to speak about is the errors resulting in the rasterizing a vector map which means you are converting a vector map to a raster map, so grid now when I say a vector map it has in a point line and a polygon. For example, I have taken a polygon here. This is a triangle that I have considered. So if I have considered this is the polygon that let us consider.

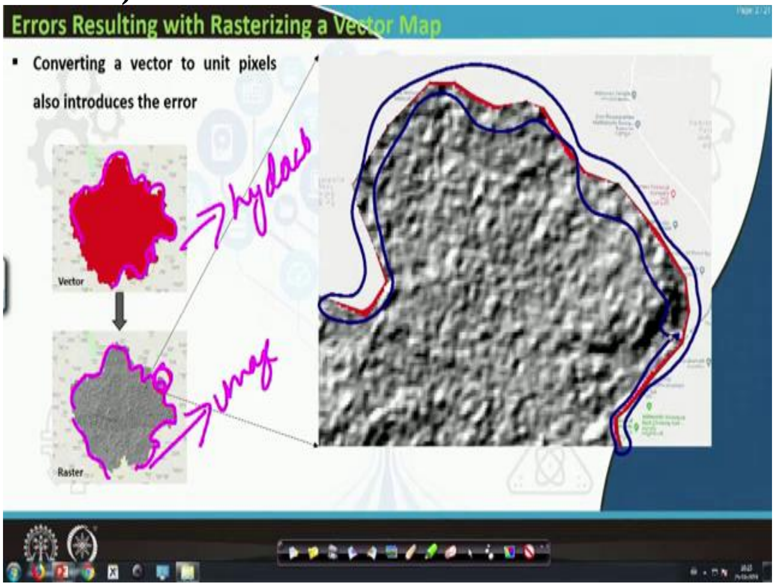
So, when we are let us say we are converting it to a raster unit which means we are converting it in a form of pixel data equal number of equal pixel size that is how we will be doing now, for example when we are looking at this area of a triangle area of a triangle identically should be about 7 units. So, when we are converting it into raster for example there have shown here how it has been converted to raster.

So, you have this as 1 pixel unit this has 1 pixel unit this has 1 pixel unit, but when you look at this polygon it is something like this. So, you are actually approximating that particular polygon to over that pixel right. So, when we are looking at any area we may consider it as 6 or 7 units depending on how the cells and sites are counted in this particular triangle. But, when you look at hypotenuse you may have 7 cell sites long if 4 cells are considered as an approximate of the diagonal.

If you consider 4 this is a diagonal, if it is approximate of a diagonal, then you will consider it as a 7 cell side long if it is 4 cells otherwise if you consider it as 3 cells, then you may look at least 6 cell side. So, this is what is approximation error that you may get in a raster map. So, if you consider very large grids, large size grids, only then your accuracy need to be need to be understood in a bigger way. If it has let us say you have the grid size extremely small, then such errors may not actually creep in if they are if the grids are like this.

But if the grid is something like this. I will represent in a different colour, if the grid is something like this. So this representation may create an issue. So, such errors also has to be handled when you are converting a vector data into a raster data. So, these are these one of those errors where which actually creeps in when you are trying to do analysis that has both vector and raster format. So, basically when you are converting this asset a raster data.

(Refer Slide Time: 05:58)



Errors Resulting with Rasterizing a Vector Map

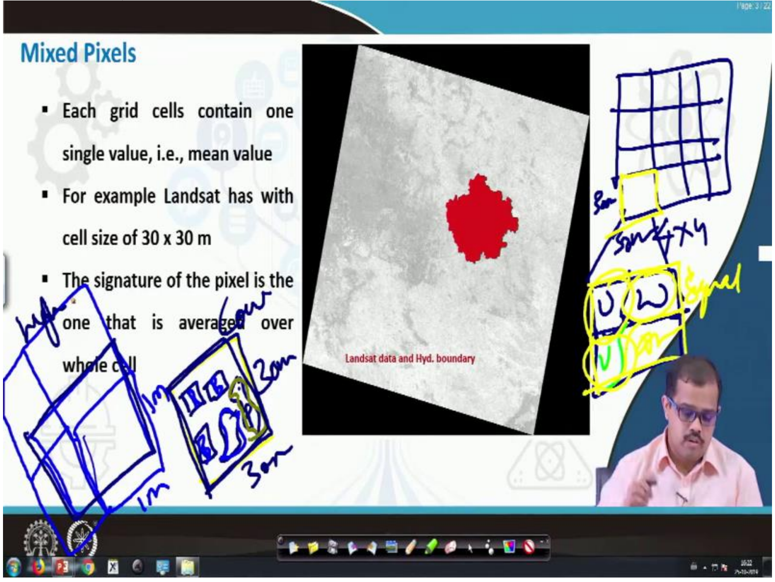
- Converting a vector to unit pixels also introduces the error

Now, I will give you another example. For example, I have considered here this is a map this is a Hyderabad boundary. Now, I have considered image that is exactly cut in terms of Hyderabad boundary. Now, this is a raster and this is a vector what basically I have done this I have taken this particular polygon based on the polygon have approximated it and converted into a unit place into vector.

So, now when you look at this particular region for example, this is what I have here macro in so, when you look at this, if you see possibly these regions and you look at these regions, these regions are those where the data is either missing or data is approximated, which means these many number of unit pixels are missing in that data which means that you are you are actually removing that data that may be useful in certain terms.

So, this is also one kind of an error that may creep in when you are actually converting a vector data into a raster data.

(Refer Slide Time: 07:24)

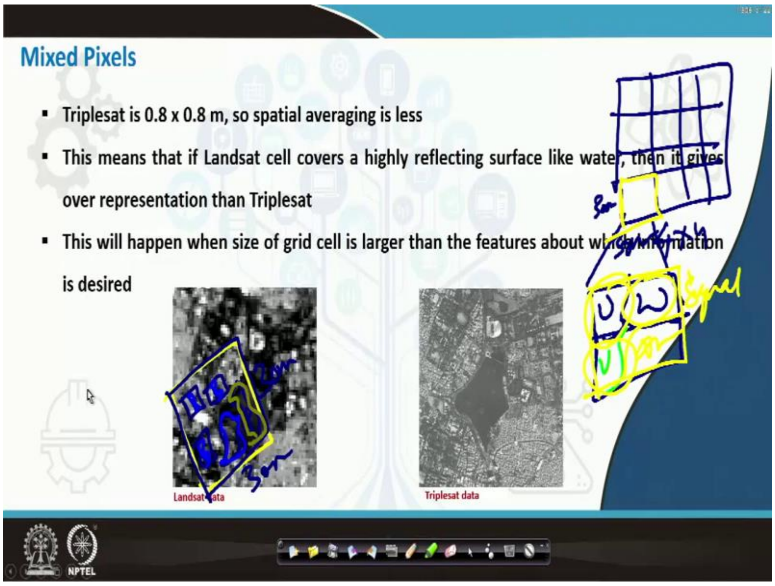


Mixed Pixels

- Each grid cells contain one single value, i.e., mean value
- For example Landsat has with cell size of 30 x 30 m
- The signature of the pixel is the one that is averaged over whole cell

So, then the second problem is mixed pixels. So, each grid cell contains 1 single value when whenever you have for example,

(Refer Slide Time: 07:34)



Mixed Pixels

- Triplesat is 0.8 x 0.8 m, so spatial averaging is less
- This means that if Landsat cell covers a highly reflecting surface like water, then it gives over representation than Triplesat
- This will happen when size of grid cell is larger than the features about which information is desired

If you take if you look at the Landsat data each pics here when you look at this as a Landsat data let us concerned certain number of rows and columns, I have just taken 4 rows into 4 columns. In case if I am constrained this each is 30 meter by 30 meter. So, if you look at your city region, a city region, if there is a 30 meter by 30 meter pixel, it means this particular pixel can have various land users here.

This part may be urban, this may be for example, water and this may be vegetation and this may be in others category right. So, now, what happens when you are actually this particular cell after imaging is being converted from analogue signal into a digital signal where it is sampled this entire region will be just converted as 1 signature this particular signature actually belongs to this particular class or a digital number.

That belongs to a certain class. Now, if this happens, this is called as a mixed pixel issue now, because of number of classes that you have, which are true representation of the real word, but when you look at your data model, you have mixed pixel data. This is one of the big issues that we face when we have a data which is of course resolution. So, when I say course solution.

These are the data which when I when I refer to course again course when I say course resolution, it is reference to something I am referring a 30 meter data as course resolution with reference to 1 meter data or 2 meter data that we have. So, when I say 30 meter each pixel will have 30 meter by 30 meter information. If I say, if the spatial resolution is 1 meter, then each pixel will have 1 meter by 1 meter information.

Which means it is giving me more details, let us say if this is your pixel, I am taking just 1 pixel out of here. So if we consider this as a pixel. Now, what happens if we consider this pixel if it is let us say a 30 meter by 30 meter pixel? Correct. So, if there is, let us say 1 building, there is 1 more building here. There is 1 more building here. There are may a small lake.

Maybe a very small water body or a lake or even a maybe a well then there may be a playground this is our let us say a playground I will put this in maybe in different colour this is a playground and maybe this particular region is water body other than this these are buildings. Now, if this is a 30 meter pixel 30 meter by 30 meter pixel, then entire region will be sampled as just 1 different class.

Whereas, when you when you are looking at other regions for example, let us say is the same thing if I applied to and 1 meter by 1 meter data, this is a 30 meter by 30 meter data, the same pixel I have it here, but it is a 1 meter by 1 meter data from different satellite, let me consider it to be one of the triple set satellite which is 80 centimetre data resampled it to 1 meter. Now, I have 1 meter by one meter data. So, this 1 meter data for example, this particular building may be falling in may be 3 4 5 pixels of this particular page.

This may be another pixel, this may be another pixel this may be another pixel. So, when you are looking at this particular image, this is a building that you are capturing in both of this image but in this is just a part of a pixel, but here it is a larger part of many pixels. So, what happens in this, this is actually a gross information or it gives you very less details about all of the quantities that are here.

But this gives you a very detailed look about what quantities are there in that this particular region, which means you say this has a better resolution, higher resolution and the one which you are referring here is a course resolution, this is a course resolution data whereas this is a high resolution data. So, that is what we have to understand when we are trying to understand high resolution course solution.

Always when you are saying some data is high resolution, please refer to what reference you are saying. For me, it the high resolution may be a 30 centimetre data or 20 centimetre data for someone who does not have an access to data the 1 meter of 5 meter data maybe a very high resolution that he or she can get some people 30 meter maybe the very high resolution. But compared to what compared to 500 meter 1 kilometre 30 meters a very high resolution.

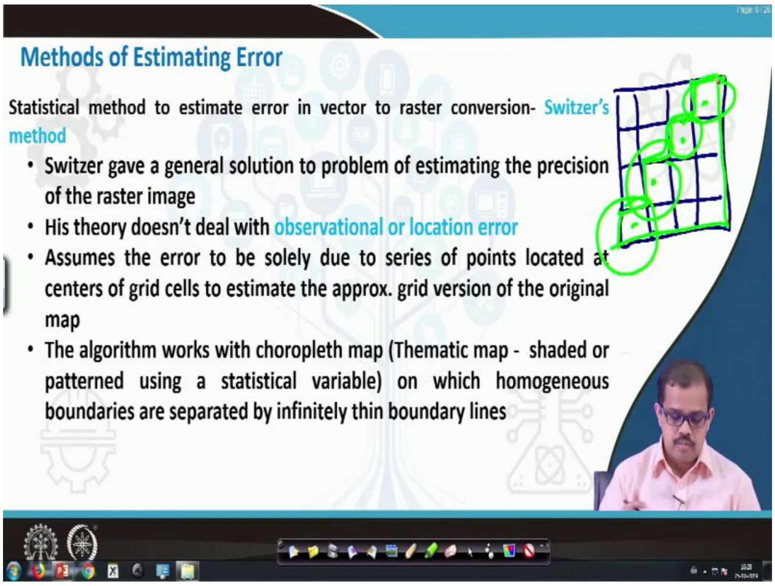
That he or she may have access 2 or 5 meter is very high resolution when compared to 100 meter or 500 meter data that he or she may have access to. So, always look at the comparative statements when you are referring to many of the resolutions are when we are referring to the pixel size etc. So always be careful when you are mentioning it is a high resolution versus a course solution or high detail versus a low detail maps.

Ok, so let us go back to what I was speaking about. So, in a medium resolution of course is not sure when compared to 1 meter data, maybe a 30 meter data what I am considering here a Landsat. So, this particular pixel will have a number of different entities into it. So these entities will create a mixed pixels issue. So when have a mixed pixel issue. Let me give you 1 more example. This particular image that you are seeing here, this is an image from a triplesat data, but this is an image from a landsat data both are of the same region.

This is probably from Bangalore. So this is a triplesat data this is the Bangalore data that we that we can see. For example, when we go look at this particular data for example, you are looking at this data here. If you zoom in, you can see that the amount of details for example, this particular thing is a lake, this particular thing, whatever you are seeing here is a water body or a lake. So, and this is the region where which has a road, these are the buildings that you can see.

But this is probably the sankey tank and this is the road that you see and this is the indian institute of science Bangalore. So, when you look at these buildings, you can very clearly get information about these buildings. So which means the detail is very high. So this is a high resolution map. Similarly, when I look at the same image here, if I look at the same image in the Landsat, if you look at this particular region which,

(Refer Slide Time: 15:28)



Methods of Estimating Error

Statistical method to estimate error in vector to raster conversion- Switzer's method

- Switzer gave a general solution to problem of estimating the precision of the raster image
- His theory doesn't deal with observational or location error
- Assumes the error to be solely due to series of points located at centers of grid cells to estimate the approx. grid version of the original map
- The algorithm works with choropleth map (Thematic map - shaded or patterned using a statistical variable) on which homogeneous boundaries are separated by infinitely thin boundary lines

I am pointing out, so, the same image is shown in a very different way, when you are trying to look at if you cannot even identify what is a boundary of this particular water body see probably this is approximate boundary of this water body, whereas, here you could have drawn the exact boundary with exact details of this water body, right. The one thing that I have done on both the images is nothing called, nothing but called as a onscreen digitizing.

I have digitized that particularly, you have to be extremely careful it is not so, direct as what I did, but you have to be extremely careful on those curves. So, when you look at this, you could not you cannot even find out where is the road whereas the vegetation which you can find out here, and where are you this is the road that you normally find out so, very well the sankey tank on the sankey road.

So, that is what is the difference between a good resolution or high resolution data and of course resolution data or even Landsat is considered to be a medium resolution data. So, when you why I try to explain this is stat for you to understand why mixed pixel issue is extremely important in order to address it is that thing that you have to understand. So, if the cell grid is extremely larger than the features about which information is desired may not be extremely visible in terms when you have a low resolution or of course resolution data.

When you have a high resolution data this is extremely accessible. So, how do we how do you find out the errors that may have crept in such vector to raster conversion? There are various different ways of calculating this error. The one of one of the very well-known method says Switzer method. So Switzer method is normally in a very approximation method. So what it basically does is it Switzer gave a general solution for estimate, like the precision of an image.

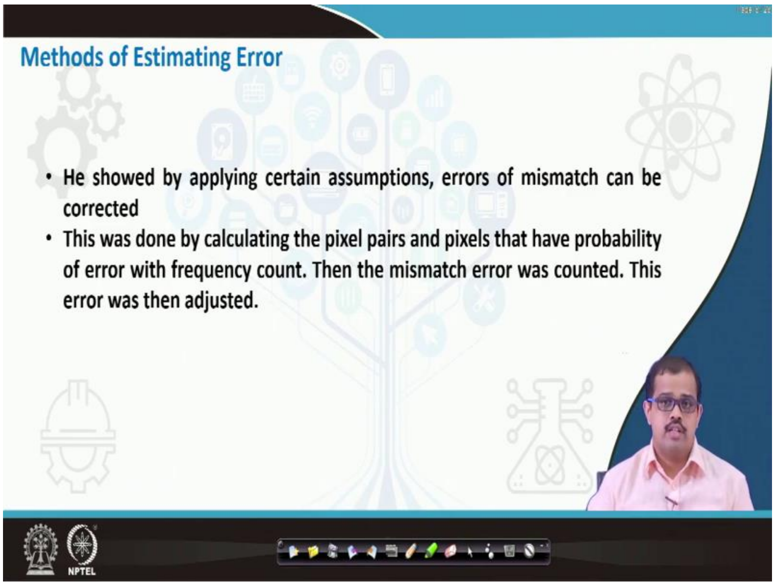
He did not given any method for estimating accuracy, but he defined it as a precision of an image, it means his theory does not deal with any observational or locational error. The very important when you are trying to understand this theory does not now think about any of the observational locational error. It assumes that error is solely based on points located at the centres of a grid cells to estimate the approximate grid version of the original map.

What I mean to say here is, if this is a particular grid that has been converted from a raster to a vector, let us say, this is a grid that I have, now I have converted a raster to vector, green represents a raster to a vector. Now, what it actually means is that it does not consider this particular information of position or the location where it is actually located in entire grid but it considers only the approximate the center values here.

And looks at how good it is whether it is displaced or it has in the estimated location that is correct or not. So, when you are looking at Switzer method that it has more an approximate method than being an extremely viable method in terms of assessment of the error. So, algorithm basically works on thematic maps for which are also called as clara park maps. Where in when I say glorified maps is a special maps which have a shaded or a pattern regions using a certain statistical techniques or statistical variable.

So on which the homogeneous boundaries are separated by very infinitely thin boundaries. So, that is why that is what they mean by it these are the homogeneously supply in infinitely thin boundaries. So, this is how the Switzer method works, it is more an approximation method. And more looking at the central weight of a pixel in order to estimate what is the error that may have encrypted. So, there are several different methods that are being used.

(Refer Slide Time: 20:04)



Methods of Estimating Error

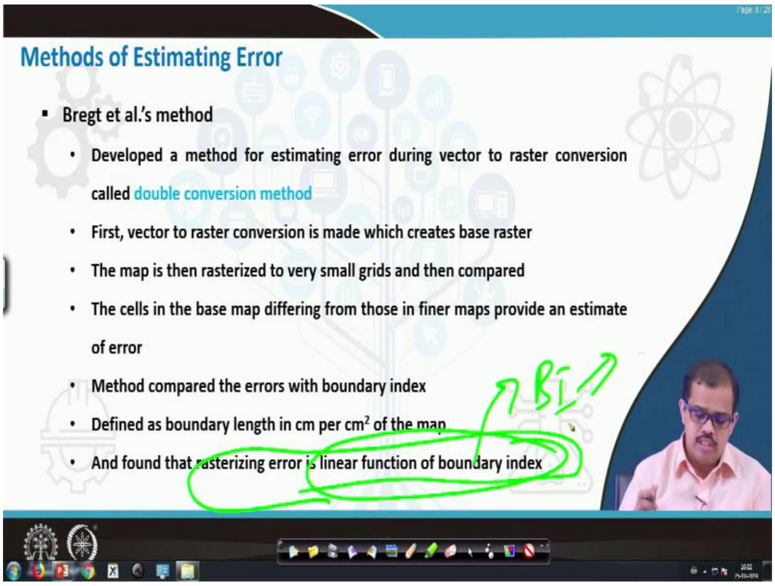
- He showed by applying certain assumptions, errors of mismatch can be corrected
- This was done by calculating the pixel pairs and pixels that have probability of error with frequency count. Then the mismatch error was counted. This error was then adjusted.

But Switzer is a very well-known method. So, here he showed this by a applying certain assumptions, errors of mismatch can be corrected. This particular algorithm had huge number of assumptions. For example, first thing is the this algorithm calculates for per polygon and then set of polygons what is the kind of error that may creep in. So, once it looks at the set of polygons it tries to differentiate between what is a mismatch.

So, once it finds out the mismatch between a pair of a polygon on a single polygon then it says that this is the amount of error that has crept in that is what is the algorithm of Switzer. So, when you are looking at the pixel pairs and pixel having that particular probability and looking at the frequency count he just estimated. But what is the major thing is he missed out at the locational information that is necessary for any GIS data without trying it to the location information.

Now, any method can fail in terms of providing the error details in a much larger context. If you have much bigger vector layers, then it is extremely difficult for us to find out the error

(Refer Slide Time: 21:22)



Methods of Estimating Error

- Bregt et al.'s method
  - Developed a method for estimating error during vector to raster conversion called double conversion method
  - First, vector to raster conversion is made which creates base raster
  - The map is then rasterized to very small grids and then compared
  - The cells in the base map differing from those in finer maps provide an estimate of error
  - Method compared the errors with boundary index
  - Defined as boundary length in cm per cm^2 of the map
  - And found that rasterizing error is linear function of boundary index

There is another method which was developed by Bregt et al. So, this was based basically on a double conversion method. So, it is also very famously called as dcm method. So, the first vector to raster conversion here was made as is it creates a base vector. So here it does not just directly give you a vector to raster data, but what basically this method does is it first gives you a base vector.

Once you have the base vector and then the map is fast rising to very small grid and then compare thus the base vector is then rasterized. So, whatever you are base raster is first created, then you have a vector map that is converted into a raster map. Once you have a raster map, each of this raster map is then converted into small grids as small as possible and then compared with the base raster.

So, now, with when we are comparing with the base base raster, it the cells in the base map differing from those in the final map that is in the raster map whatever is the cells that exist and are defining with the small gridded cells that you have converted from a vector to a raster this provide an estimate of a error the amount of error is then defined by the thin difference. That may be there in the layer that is converted from vector to raster based raster map.

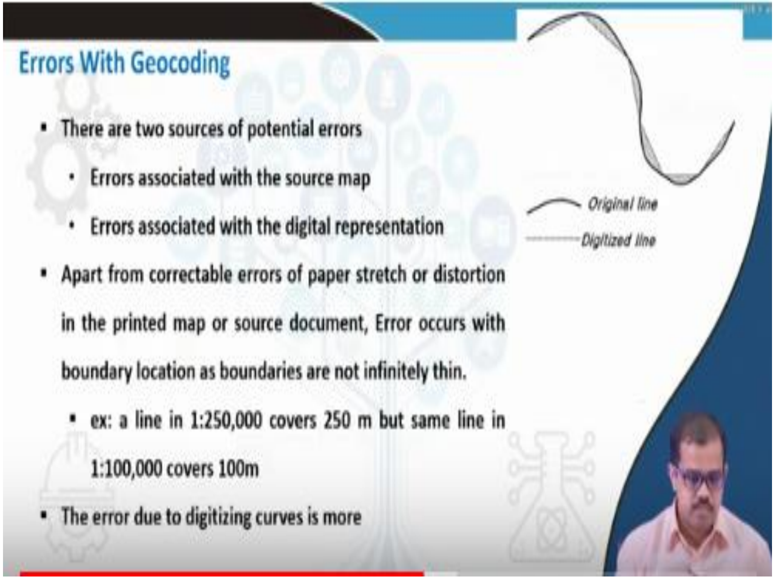
Then this method can basically looks at the locational error and also looks at the errors and the boundary index that is why it is extremely efficient in terms of conversion from a vector to raster. So, most importantly it defines the boundary length in centimetre per centimetre square off of that particular map. Then one thing that you can find out here is this rasterizing error is linear function of the boundary index that it calculates.

So, if you can under a few understand what is a linear function of the boundary index, how do you found to find a boundary index, So, if you calculate the linear function of this and try to substitute the values that are existing with each of the vector conversions. You will be able to find out how do you estimate an error between the thin lines of base raster and vector to raster converted data.

So, this one method. So, most of your software that you may use in future course of time any of the software let us say arc GIS QGS etc they have either one of these methods or most more evolved methods that for example mcm etc. So, these methods are then you used in order to do a vector to raster conversion, so, you get more smoother polygons, smoother rasters than what you may have seen in this examples.

So, more and most of the GIS software today you do not understand what is the background of it, but these are the algorithms that actually drive those tools. These are the science behind those tools, which are driving those software in order to convert the vector to raster.

(Refer Slide Time: 24:50)



Errors With Geocoding

- There are two sources of potential errors
  - Errors associated with the source map
  - Errors associated with the digital representation
- Apart from correctable errors of paper stretch or distortion in the printed map or source document, Error occurs with boundary location as boundaries are not infinitely thin.
  - ex: a line in 1:250,000 covers 250 m but same line in 1:100,000 covers 100m
- The error due to digitizing curves is more

So, there are certain errors with geocoding. For example, when you look at 2 potential source, 2 major potential source that you can see now, any of the images that are geocoded it has error associated with the source map. First of all, a source map if it does not have I mean, it has to be extremely precise, extremely precise and accurate. If there are the accuracy and precision is much lesser.

Then obviously the error is there in the source map by itself and errors also as are associated with a digital representation that this these errors are quite easy to understand when you are looking at your own data set. So, apart from the correctable errors of the paper, stretch or destruction, up errors also occurs with the boundary location or the boundaries that are in not infinitely thin. For example, I gave an example of how the vector is converted to a raster.

Now, let us say I have a boundary line just to represent it as a form of a map. Now, I give a boundary line which may be 1.0 as thin or 10.0 as thin depending on the way it has to be represented by me. Now, if I have 10.0 as thin. So it may cover that entire pixel which is actually converted into a raster map, whereas 1.0 may even give some a difference between the conversion that is already there. Pink lines would have been seen.

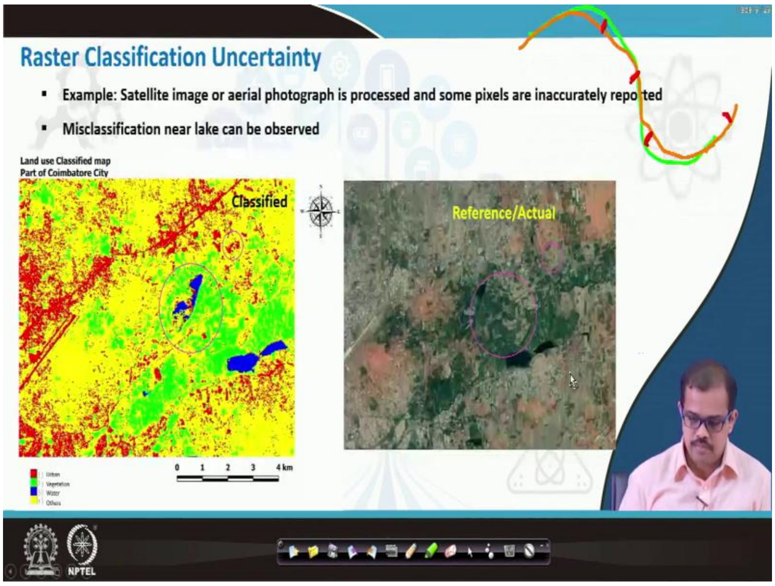
So, that gives more of a different kind of representation. So, again this if this is digitized and then converted then it may also create an issue for example, at 10.09 may lie in between or to one of the sites, so, that again creates these digitizable errors. So, these are different errors that you can see when you are trying to geocode a line geocode and image. For example, a line in 1:250,000 covers about 250 meters.

But the same line in 1:100,000 covers 100 meter. So, for example, if you are drawing a line the same in a 2 different scale map, so 250 meter verses 100 meter, that is what is the difference that you may get in the boundary line conditions and there is as I explained the digitizing curve errors, so, there are huge number of errors for example, the one image that I have shown here if this is a particular line the exact line that I have

Ok, then this we have seen more often if I let us say I start digitizing this line. So, what I do is I may because of the my digitizing if I look at the digitizing I may just want so, which means this is an error that we have created and again here is the error and this is error. So, these are the digitization errors that we would have done for example here. So, these also can creep in when we are digitizing, so, errors can be everywhere.

Only thing is that it is up to the user it is up to the person who is actually creating the map to look at the standardized methods of looking at it on look at in a more detailed way in order to correct those different errors that may be in the entire data set.

(Refer Slide Time: 28:24)



Raster Classification Uncertainty

- Example: Satellite image or aerial photograph is processed and some pixels are inaccurately repo ted
- Misclassification near lake can be observed

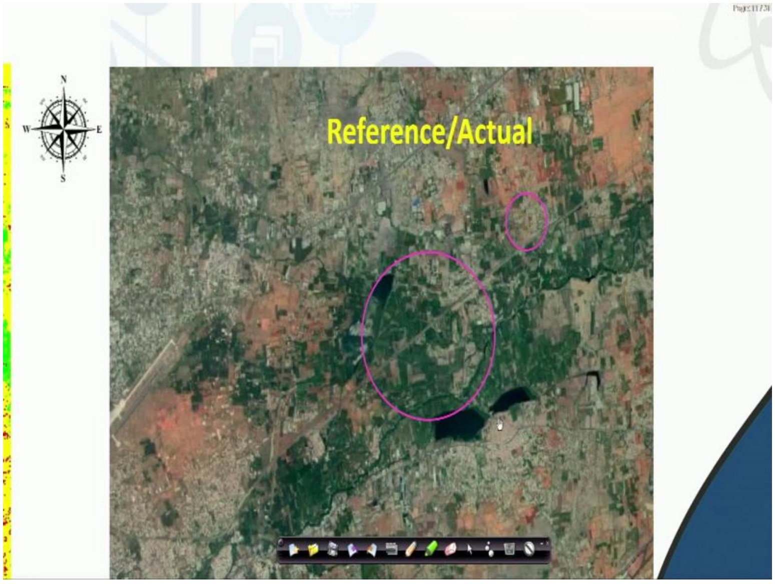
So, when we look at raster classification, when I say raster classification, I am actually looking at the land use classification. For example, when I look at this particular on on the left hand side, you find a satellite data this is a Landsat data, this is representing a particular region I think this is a Coimbatore city. So, this particular classified map is we have created a land use map.

Here where you have classes as urban vegetation water and others red this representing urban vegetation is represented by green and water is represented by blue, others is represented by yellow and we have used an algorithm that is called as gaussian maximum likelihood classifier. So, now, when you detailed look at this particular map, for example, because of your classification, the input signatures that have been given.

Maybe you have given a huge number of heterogeneous signatures for this to be classified this image to be classified as a land use map. Now, what has happened is that you can see, there are certain errors that has already creped in inform of misclassification errors you can see this thread here, around the water body, you can see this part here, where you have a huge amount of red. I can zoom it in and show it to you probably. So if we zoom it into this region, you can see this particular region of the water body is completely misclassified.

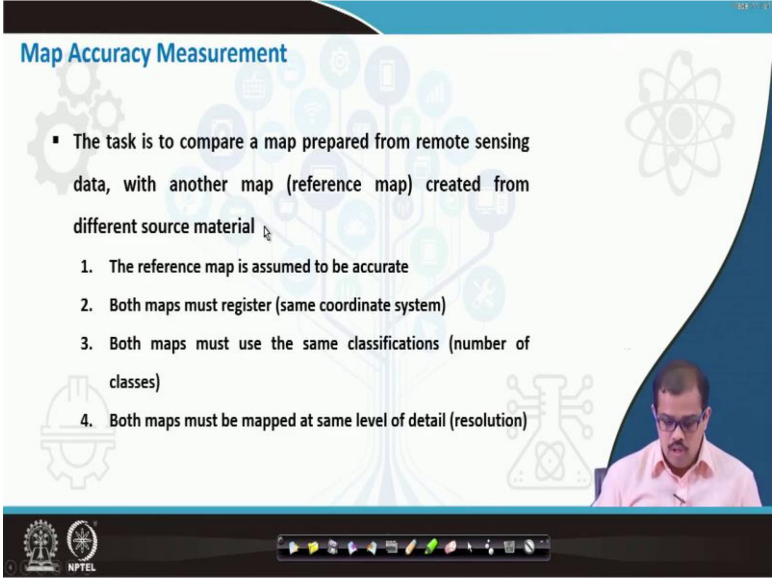
So this is a region that is misclassified is a region that is misclassified and when you see this entire water bodies misclassify.

(Refer Slide Time: 30:21)



Similarly, we have another image here, this a reference image, when you look at this, this particular entire region had does not have a water at all. This particular region that is not actually not an water body, but an open space is classified as nothing but an water body. So, when you are looking at this particular water body there has a certain amount of water, but here the urban was shown, but this particular region is not urban, but you have thick vegetated region, if you look at it, it is more of a vegetation. So, when you look at such issues, the these are the issues which is actually,

(Refer Slide Time: 31:03)



Map Accuracy Measurement

- The task is to compare a map prepared from remote sensing data, with another map (reference map) created from different source material
  1. The reference map is assumed to be accurate
  2. Both maps must register (same coordinate system)
  3. Both maps must use the same classifications (number of classes)
  4. Both maps must be mapped at same level of detail (resolution)

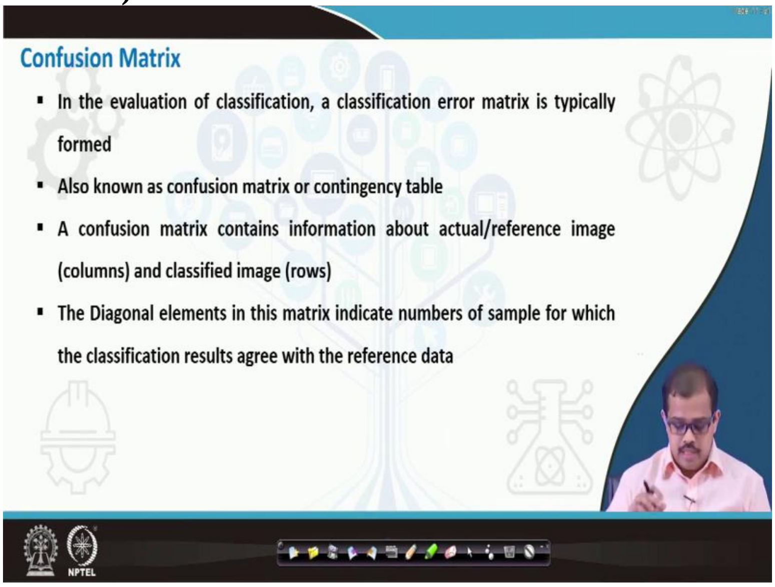
May cause certain accuracy assessment measures to be put in place. So, there is something called as a map accuracy measurements. This is the task to compare a map that is prepared from a remote sense data and another map we just created from a different source material basically used as a reference map, where it has accurately mapped onto the ground or on to the real world scenario. So, we assume the basic assumption is that the reference map is accurate.

That is a very basic assumption that we normally use and both maps when you are have I am repeating it again and again both maps when you are actually comparing both of them should be in the same reference system when I said reference, same datum, same coordinate system same projection. So, both maps must be of the same classification, if you have used 4 classes. And the first class is urban second as water third is hesitatiation and fourth is others, 4 classes of those categories, then the other maps also should be of the same 4 classes, you know.

If you have to compare about the accuracy and both maps mapped at the same level of detail. So if you are understanding this, then you will be able to understand what now what are the specifics that we will look at when you are looking at a particular map. So, everything should be same only thing is that the source the way it is generated is two different places one is it is a reference map is considered to be accurate more accurate in terms of representing the real world phenomena.

Whereas, the user created map is through algorithms that and as more of approximated map wherein we training the algorithm to automate the process of understanding the every land use parcel in the entire region. And classify every pixel into a particular class that is what we mean by classification map or a user map under reference map.

(Refer Slide Time: 33:08)



Confusion Matrix

- In the evaluation of classification, a classification error matrix is typically formed
- Also known as confusion matrix or contingency table
- A confusion matrix contains information about actual/reference image (columns) and classified image (rows)
- The Diagonal elements in this matrix indicate numbers of sample for which the classification results agree with the reference data

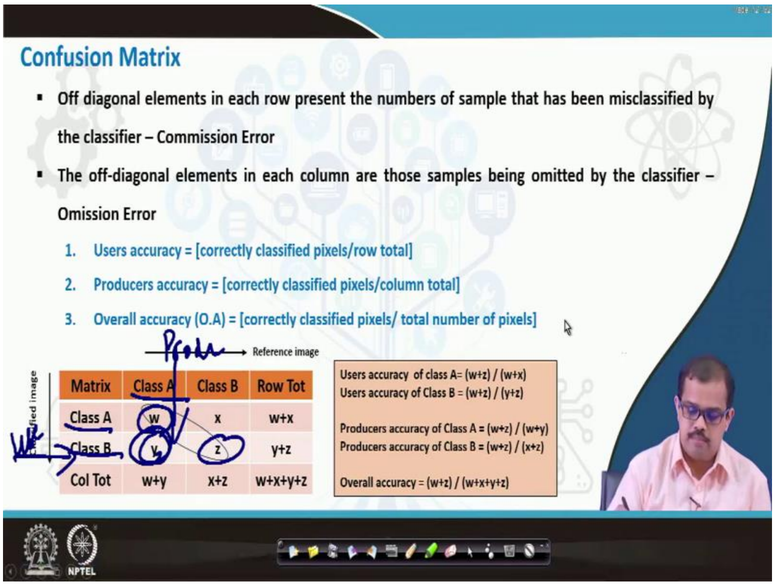
Now, so, if this is understood, then the first way of looking at this is using a confusion matrix or very well known as error matrix. In a evaluation of classification of classification and our metrics is a typically form it is a very well-known method. And most of the classification systems use it as one of the authentic ways of looking at how best is your classified data. When I say classified data, it is dividing the earth surface into different user classes.

When I say user classes, how humans have converted that a particular parcel of land into his or her own news, it may be urban, if you are looking at level two classification, if I am looking at level three classification, it means deeper classification I made a divided into industries, buildings, residential buildings, etc. Now, if I have to really go into much deeper sections, I will divide it into residential building as SIG, MIG, LIG etc.

So, the this the amount of classification detail that you need depends on the user. So i it is not upto anyone to say this is what the classification detail has to be what is the thing that you are trying to understand the phenomena you try to look at that amount of classification detail and one of the classification details. That I would give an example here is the basic classification of level 2 level 2 class. So, now, if I have created 4 classes let us say urban, water, vegetation others.

So, you will have 2 images with the same reference details. Coordinate system is same datum is same and a projection details are same. One is a user map which you have created using your field data collection and also with your no one is user map.Which you have collected in a physical field data collection with signature collection and over the heterogeneous region whereas other one is a reference map which is considered to be almost accurate. Now, these are 2 maps when we put it in in the form of a diagonal elements.

(Refer Slide Time: 35:22)



Confusion Matrix

- Off diagonal elements in each row present the numbers of sample that has been misclassified by the classifier – Commission Error
- The off-diagonal elements in each column are those samples being omitted by the classifier – Omission Error

1. Users accuracy = [correctly classified pixels/row total]
2. Producers accuracy = [correctly classified pixels/column total]
3. Overall accuracy (O.A) = [correctly classified pixels/ total number of pixels]

Matrix	Class A	Class B	Row Tot
Class A	w	x	w+x
Class B	y	z	y+z
Col Tot	w+y	x+z	w+x+y+z

Users accuracy of class A= (w+z) / (w+x)
Users accuracy of Class B = (w+z) / (y+z)

Producers accuracy of Class A = (w+z) / (w+y)
Producers accuracy of Class B = (w+z) / (x+z)

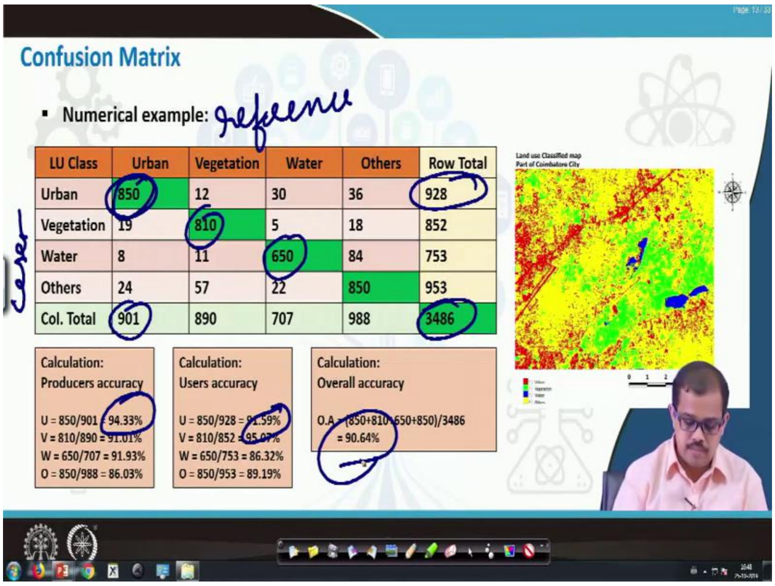
Overall accuracy = (w+z) / (w+x+y+z)

I will show you how it has to be done. For example, if we have 2 images or 2 images, this is where you have a classified image, this is where the reference image reference image lies. Now, let us say you have 2 classes in this class A and class B. So, now, this is class A, this is class B. Now, class A which was supposed to be class A in reference map is also class A in classified map that is a user map.

So that is this main number of pixels class B, which was supposed to be class B in the reference map is also class B in that user map that is z number of pixels. So the totally classified correct pixels is nothing but w + z, now a particular pixel, y is classified as for example, when you look at this particular image, this is when we call this as this is a producer accuracy.

And this is considered to be in user accuracy when you are computing this y pixel ÷ by total number of pixels? So this computes your producers accuracy of a particular class I will give you better give you an example. So you will under you will be able to understand it much better.

(Refer Slide Time: 36:51)



Confusion Matrix

- Numerical example: reference

user

LU Class	Urban	Vegetation	Water	Others	Row Total
Urban	850	12	30	36	928
Vegetation	19	810	5	18	852
Water	8	11	650	84	753
Others	24	57	22	850	953
Col. Total	901	890	707	988	3486

Calculation:
Producers accuracy

U = 850/901 = 94.33%
V = 810/890 = 91.01%
W = 650/707 = 91.93%
O = 850/988 = 86.03%

Calculation:
Users accuracy

U = 850/928 = 91.59%
V = 810/852 = 95.07%
W = 650/753 = 86.32%
O = 850/953 = 89.19%

Calculation:
Overall accuracy

O.A = (850+810+650+850)/3486
= 90.64%

For example, if I am constraint this as an example here, for example, I have 4 classes that I have taken, I have 850 pixels, which are supposed to be urban are classified as urban, I have 810 pixels we are supposed to be vegetation have classified as vegetation, I have 650 pixels we are supposed to be water are classified as water in my users map, but there are certain other pixels for example, if other than in off this diagonal elements.

All the pixels are those misclassified pixels. Now, what the first thing that I compute here is what is a producer accuracy and that is what is the accuracy of your reference map with respect to the details that are there. So, which means you say that now, this is where you take your reference map reference map is always on the columns whereas, users map is always on the rows.

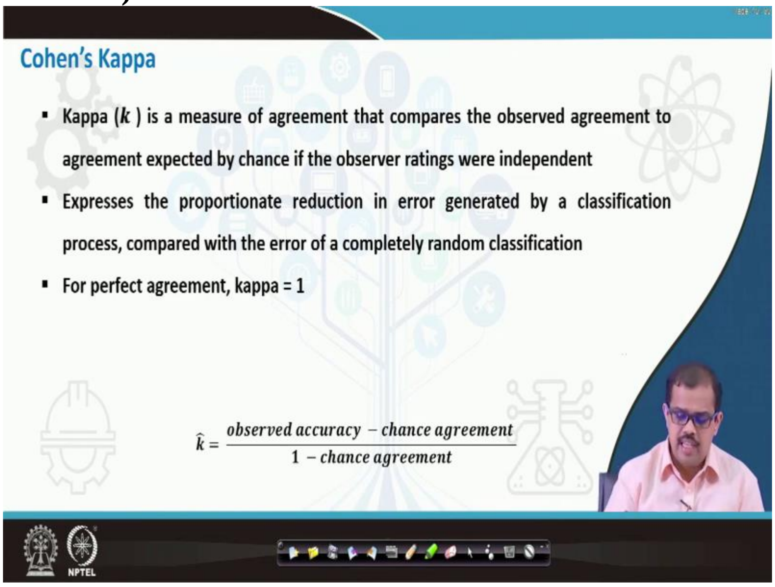
So, now, when you want to look at the producer accuracy of urban you look at 850 pixels that are perfectly classified and you have total number of pixels is 901. So, 850 by 901 is this is your producer accuracy which is relative to your reference data. Now, if I have to calculate how good is my classification accuracy, so, you look at the same 850 pixels ÷ by the row total that is 928.

So this gives you the users accuracy which is 91.58. So, why do you need a producer accuracy and the user accuracy is to estimate how good is your user classified data was the producer refer data so, comparison of this will give you extremely important details in terms of how accurate is your information. Now, if you want to calculate what is the overall accuracy of your image constraint both user accuracy one producer accuracy.

You would be constrained 850 + 810 this is 850 + 810 plus 650 + 850 ÷ by total number of pixels. Total number of pixels is given here it is including the error pixels that are there total number of pixels in that particular image. So, that gives you the overall accuracy of the this means that your data has 90.64% of the data in your classified data with reference to a data that is considered is accurate.

Without this kind of validation, it is difficult to provide any details to any governing agency or anyone saying that whatever the classification has been done, this is one of the ways of looking at the how good how accurate is your particular data. Now, we looked at it another measure is looking at, for example, another measure is to look at us using a kappa.

(Refer Slide Time: 39:55)



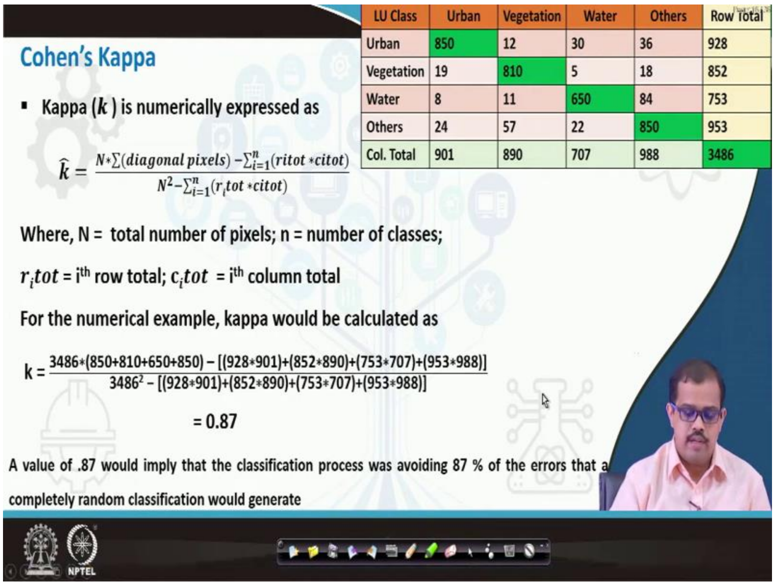
Cohen's Kappa

- Kappa (k) is a measure of agreement that compares the observed agreement to agreement expected by chance if the observer ratings were independent
- Expresses the proportionate reduction in error generated by a classification process, compared with the error of a completely random classification
- For perfect agreement, kappa = 1

$$\hat{k} = \frac{observed\ accuracy\ -chance\ agreement}{1\ -chance\ agreement}$$

So, kappa or Cohen’s kappa is a measurement of that compares the observed agreement is in agreement to expected by chance if that observed rating we are independent. It is not quite dependent in terms like your now overall accuracy. For example, when you look at kappa here. Observed accuracy minus a chance agreement divided by 1 minus the chance agreement. A chance that particular that by that pixel belongs to that particular class. So, it goes pixel wise the highest value lies between 0 and 1 if it is 1, it is perfect agreement. If it is 0 it means to say that there is absolutely no agreement between the pixels.

(Refer Slide Time: 40:44)



LU Class	Urban	Vegetation	Water	Others	Row Total
Urban	850	12	30	36	928
Vegetation	19	810	5	18	852
Water	8	11	650	84	753
Others	24	57	22	850	953
Col. Total	901	890	707	988	3486

Cohen's Kappa

- Kappa (k) is numerically expressed as

$$\hat{k} = \frac{N * \sum(diagonal\ pixels) - \sum_{i=1}^{n}(ritot * citot)}{N^2 - \sum_{i=1}^{n}(r_i tot * citot)}$$

Where, N = total number of pixels; n = number of classes;

$r_i tot$ = ith row total; $c_i tot$ = ith column total

For the numerical example, kappa would be calculated as

$$k = \frac{3486*(850+810+650+850) - [(928*901)+(852*890)+(753*707)+(953*988)]}{3486^2 - [(928*901)+(852*890)+(753*707)+(953*988)]}$$

$$= 0.87$$

A value of .87 would imply that the classification process was avoiding 87 % of the errors that a completely random classification would generate

So, let me take an example for this also so, that you will understand in a better way kappa is calculated by this particular formula that is represented here, this is a formula now, if the same table is considered for this particular analysis. So, what we this is a total number of pixels so, the total number of pixels multiplied by the correctly classified pixels that is 850, 810, 650 + 680 minus.

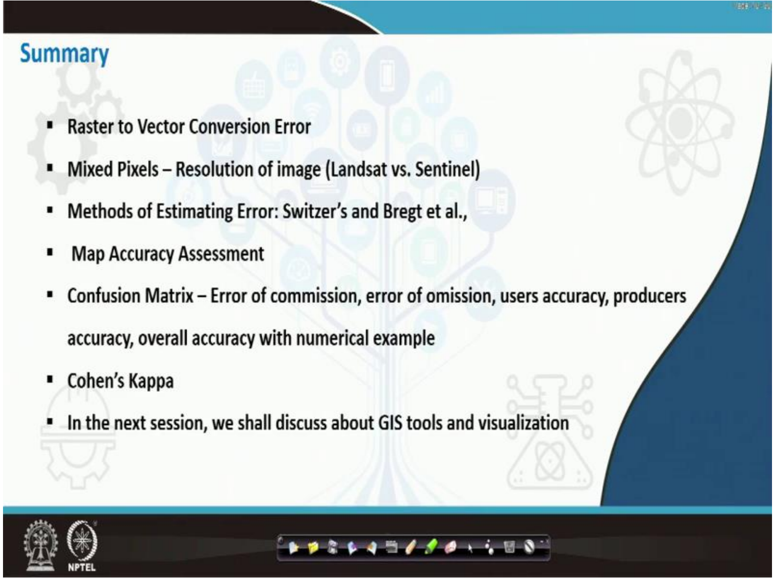
Then you have 928 into 901 which means the row total 1 into by column total 1 plus that is for urban, then growth row total for vegetation × by column total for vegetation, then row total for water into by column total for water, then row total for others multiplied by column total all of these are some, divided by the square of this total number of pixels that are there in that particular image minus again the same row total multiplied by the column total summed up.

So, this gives you the kappa value for example, let us say whatever I have considered here have got a kappa value of .87. Now, this value of .87 implies that the classification process was avoiding 87% of the errors that are completely random that completely random classification will generate. So, you are avoiding 87% of errors. What many of these students do here is when they get 0.9 kappa they say sir my accuracy is 0.9.

That is not your accuracy is not 0.9, but it means to say that a you have actually avoided the 90% of errors that any random classification or generate that said, you are not actually vary providing that this is how this is accuracy of a data. Kappa gives you what is the avoidance rate in terms of when the misclassification would happen in other random classification measures.

So, that is what the kappa is this another way of looking at the raster data which is useful classification. So, now, this is about kappa, these are different methods of looking at raster data.

(Refer Slide Time: 43:12)



Summary

- Raster to Vector Conversion Error
- Mixed Pixels – Resolution of image (Landsat vs. Sentinel)
- Methods of Estimating Error: Switzer's and Bregt et al.,
- Map Accuracy Assessment
- Confusion Matrix – Error of commission, error of omission, users accuracy, producers accuracy, overall accuracy with numerical example
- Cohen's Kappa
- In the next session, we shall discuss about GIS tools and visualization

So, now when we look at the summary we looked at how to do I mean what are the errors that may come from raster to vector classification errors that we looked at. We looked at mixed pixels I explained what is a large scale image and a small scale image and why the mixed pixel would come. So, looking at we looked at Switzer method and Bregt et al method I would recommend it k I have not spoken here mathematically.

But if you want to really understand both of these methods, you have to look at the background the mathematical background behind it, it would be very easy for you to understand it in case you understand what is it what do you mean by probability and the matrix operations. Then we looked at map accuracy, accuracy assessment, we looked at two different methods confused confusion matrix and Cohen's kappa.

So all of these put together a systems where we are trying to look at the error part of the data which is extremely important when your data is you are actually building up the data. So, in the next class, we look at how what are the different GIS tools that we have and how can we do the visualization part of it. So, as of now, I what I understand is that most of them who have taken this course would have understood how the GIS data is generated.

What are different methods and what are the errors that may creep in. So, now we are graduating towards building the GIS data the first thing we have generated, we have thought a lot looked at the generation of data, then we have looked at what are the different measures that we have to consider in order to have a good data. Now, the next part we will look at is how do we store the data, maybe in the next lecture series.

I would look at what are the different databases and how it is stored. So, that is how we graduate towards the end of the course where the software would be thought. So then we will next in next class, we will look at GIS tools and visualization as a part of just building up for the database analysis. So let us meet in the next class. Thank you very much. See you.