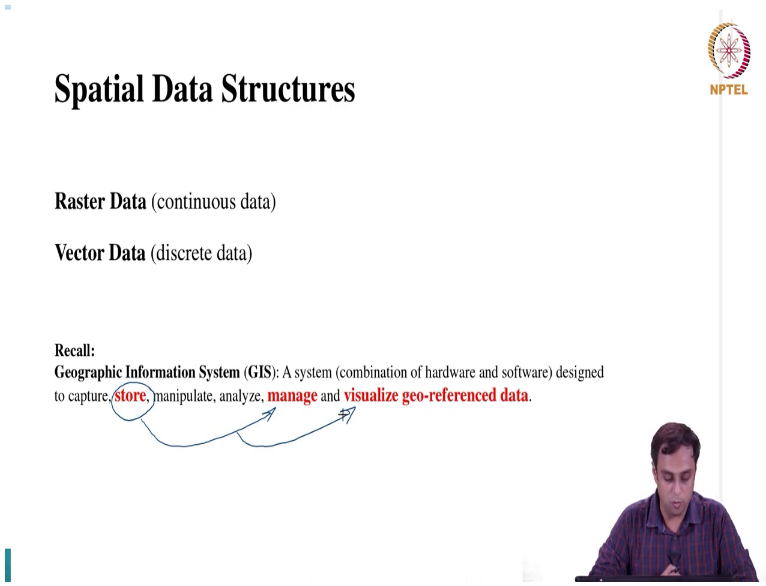
Spatial Statistics and Spatial Econometrics
Prof. Gaurav Arora
Department of Social Science and Humanities
Indraprastha Institute of Information Technology, Delhi

Lecture - 03
Spatial Data Structures

Hello everyone, welcome to the 3rd lecture of Spatial Statistics and Spatial Econometrics. My name is Gaurav Arora.

(Refer Slide Time: 00:31)



And, we will today study what are called as Spatial Data Structures. So, up until now we have seen some prominent examples of where spatial statistics is useful. We talked about how this dimension of location in determining statistics, this query of where added to how and what in traditional statistics brings to the table some really new analysis of how the data can be seen, what can be derived from the data, what can we learn from it. And, we saw that you know the economic survey of India recently you know went ahead and used this dimension of location and understanding development you know parameters in India. We also looked at you know two components of spatial data that is spatial heterogeneity and spatial dependence.

So, spatial heterogeneity is more a more of a large-scale global phenomenon whereas, spatial dependence as we saw in the previous lecture is a more local phenomenon where values nearby in space might be behaving similarly to other values in their locality compared to values that are farther away. So, today we are going to sort of follow up on what we had seen towards the end of you know the last lecture which is this concept of remote sensing.

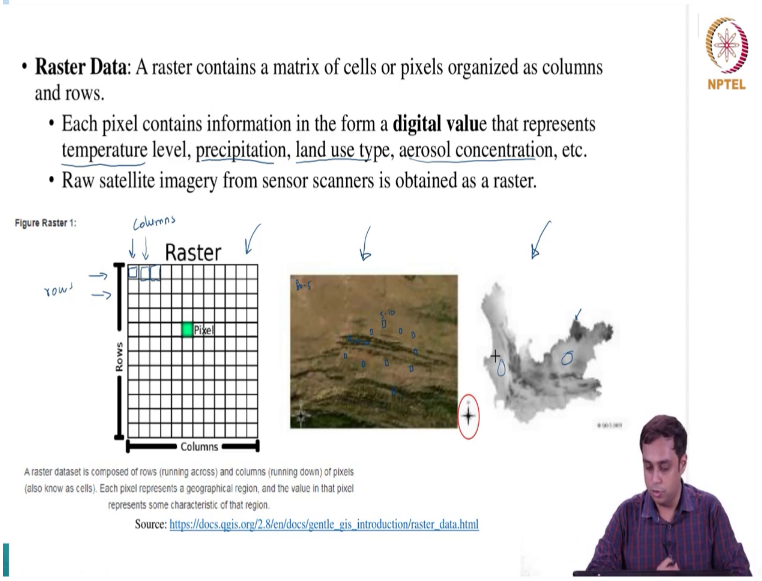
That is we are trying to where are these spatial data coming from you know what is the data generating process and what we looked at was that there are these remote sensors placed on satellites that are going around throughout the year over, you know, over the over each location on earth. And, they are basically transmitting back data that we are then able to you know store, manipulate, analyze and then visualize for our purposes.

But, then when we actually have to store, manage or visualize these data you know we need data structures right. So, we need a structure to store these data and that is the topic of today's lecture. So, I have on your screen two different types of data structures: one is called as the raster data, the second is called as the vector data. I am also characterizing on this slide that the raster data are continuous data and we will talk about what this really means and vector data are discrete data.

This is the fundamental distinct fundamentally distinct attribute between these two data sets that is you know the raster data are continuous in nature and vector data are discrete in nature in the way they are stored, managed and visualized. And, I want you to recall you know the definition that we sort of cave at the time when we talked about the geographic information system.

And, we said that this is a system which is designed to capture, store, manipulate, analyze, manage and visualize geo reference data. So, data structures start right at the place where we have to store data and then they have a fundamentally important role to play in how the data are managed and then you know how we can present or visualize this data you know in our own work.

(Refer Slide Time: 03:53)



So, let us get started. So, the first item is the raster data. A raster data contains of a matrix of cells or pixels organized as columns or rows and there is a picture right in front of you where you can see that you know the raster data are organized as rows and columns. So, you have rows, you have rows and you have columns vertically. So, right the rows are organized, you know, horizontally and the columns are organized vertically right.

And, what happens is that for these rows and columns you know you have these little cells that you can see on your screen, the cells that you can see on your screen. Each cell stores a unique digital value ok. And, this digital value when put together in a matrix of rows and columns allows us to visualize, ok. So, each pixel contains information, right, in form of a digital value.

So, now we are quantifying information from an image to a measure right and this digital value can represent multiple things. I mean almost anything you can think of which can be sensed remotely or it can be sort of, you know, put on a image which is spatially delineated. So, I have examples here you know you can have digital values that are temperatures.

So, it could be precipitation levels, it could be land use type. We have seen many examples of land use, type till now we had urban land use, agricultural land use, forest land use and for each of these land uses imagine if you had values 1, 2, 3, 4, 5, 6 and so on for different-different types of land uses if you assign those values to different parts of this matrix in front of you, you can find a land used profile.

For example, the Delhi NCR or the Chennai metropolitan region or Mumbai metropolitan region and so on and so forth. We also have aerosol concentration as an example. So, pollution, you know, that is another matter that we have discussed a bit about earlier, you know those can also be stored as unique digital values in each pixel in these data right.

Now, raster data are the most fundamental form of, you know, spatial data, satellite imagery that is coming from sensors are raster data right. If you take a picture on your phone from your own camera, if you display it on a computer screen and you if you keep on zooming, the most you know fundamental or units that unbreakable units that you come across are pixels right; that is the highest resolution at which you can look at that image.

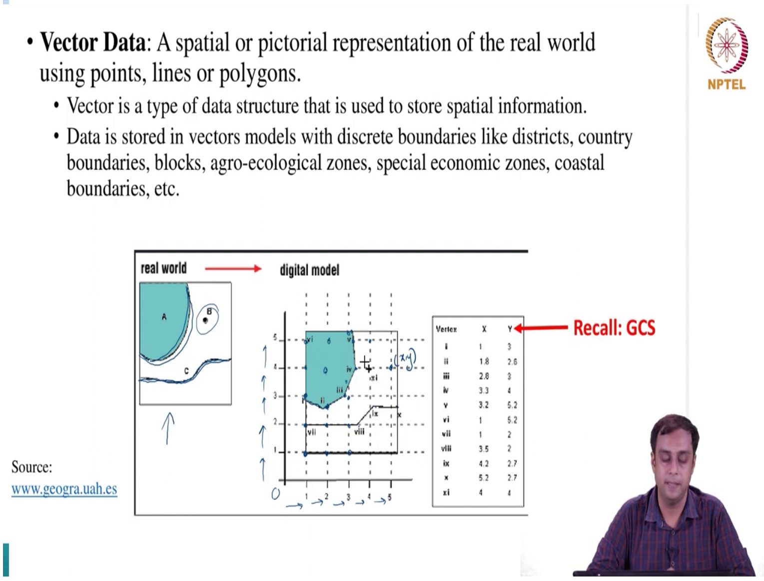
And, each pixel has just one color, that color represents really a digital value right. So, raster data is something that we come across every day, all the time you know as we you know look at you know how the world is captured on cameras of different kinds. So, in the example here if you look at the image you know in the middle that is a raster image right.

What is happening is that you know you have if you keep zooming in, you will find the smallest units that from where you know you have a uniform understanding of each location will be a pixel, each pixel will have a value, right. The value will represent here in this case in this middle picture, middle image will be a color. For lighter colors you will have some values, you will have let us say values that will vary from 0 to 5.

For slightly darker colors you may have values which varies from 5 to 10 and then from for these colors where you have darker formations, you might have values ranging from 10 to 20 right. So, and this you know scattered across space provides us a pictorial understanding of the real world. The third image is also quite similar. So, you have darker regions which can have you know higher values or you know even you know the values can be reversed.

So, the darker regions can have lower values, the lighter regions can have higher values and they can represent different things. In the last image what you are looking at is elevation on ground right. So, darker images are where you have higher elevations. So, you have ridges and hills you know on ground and the lighter regions are where you have plains right. So, they are all stored as unique digital values coming together to be seen as a raster image.

(Refer Slide Time: 08:48)



The second form of spatial data structure is called as the vector data. So, vector data is a spatial or pictorial representation of the real-world using points, lines and polygons. So, now you are basically taking the original you know pictorial image of the real-world you know scenario that is of interest and then you are converting it into structures like lines, polygons and points. And when you visualize those in that format you know what you get is a vector data. So, vector data; obviously, is a data structure that is used to store spatial information that we have talked about earlier.

And, the data is stored in vector models with discrete boundaries like districts, country boundaries, blocks, agro-ecological zones, special economic zones, coastal boundaries and so on and so forth right. So, if you are storing data in you know predetermined administrative lattices; for example, districts in India or at taluk level or at village level then the kind of data set that we look at is called as the vector data right.

Here you do not have you know information merely stored in, you know, square cells that are delineated by rows and columns alright, you have irregular boundaries which could be coastal regions which could be different different you know districts and so on and so forth. But, this type of storage of information is useful because ultimately when for example, for public policy for governance, you know, usually, the policies are sort of implemented at the district level or at the taluk level or at the village level or a local municipality level.

And so, if you want to study spatial impacts that is the location, you know, location understanding, location wise understanding of how policies let us say education policies, schools impact welfare of local communities. Then we better, you know, learn how to study how to sort of visualize vector data or to work with data which is in vector format.

So, the image here that you look at is converting a real world you know data structure which is let us say a raster data structure on your left which is nothing but an image taken from a phone from a satellite sensor you know something. So, it is a remotely sensed image. It has a region A, a region B and a region C. What it looks to me is that region A seems to be some kind of an agricultural area.

Region C which is a line looks like a road right and B which is a larger dot maybe looks like a pond right. So, there are these three features in this region that we have received in the original real world data format and now we are going to convert it to a vector data format. So, what is being done you can see that first of all you have a coordinate system right.

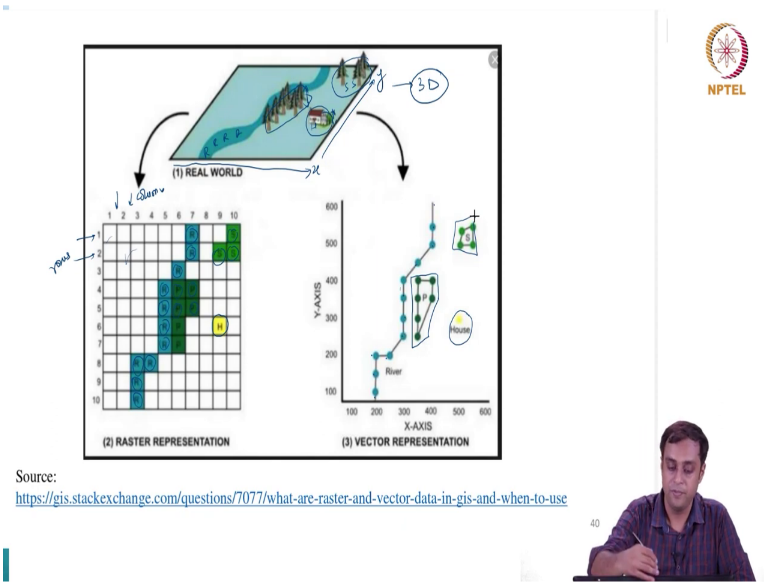
So, you have an origin and then you have a coordinate system where you have the x axis sort of. you know. broken down into discrete steps. So, discrete steps of 1 to 2 to 3 to 4 and 5 and the vertical axis sort of that is the y axis is also similarly broken down into discrete steps right like we said vector data are discrete data. So, we have these discrete steps in which you know the x, y the two-dimensional space is broken down right.

When we do that, we get these points, when we join you know all the sort of grid cells that we have created as discrete steps we get these points on you know on the x y plane right. And, we are used to sort of understanding each point will have a x comma y coordinate right. And, if you recall when we look at these coordinates you know we are basically using you know the GCS system of creating geo referencing and the creating a geo coordinate system right.

Something that we have studied earlier, I would sort of motivate you to go back and look at it to understand how these coordinates will come about for a real-world data set on land right. Once we create these points what we do is then we collect, we sort of discretize, we discretize the space that we have in circular form into discrete points which touch upon different regions in you know in my study area.

So, I get my so, my circular sort of semi-circle, you know, representation of region A is now a polygon with 1, 2, 3, 4, 5, 6, 7 vertices, right. Similarly, this road which was a which was a curvy road, when it comes to the data vector data format it is a line with 1, 2, 3 and 4 vertices right. And, that pond B simply reduces to a point coordinate B right, that point coordinate which will be basically the center of mass of this point B ok.

(Refer Slide Time: 14:24)



So, let us move forward and now what I am showing you is a contrast of a real world when you know stored as a raster image or when stored as you know a vector data or a vector image right. So, what you see, first of all, now the real world is in three-dimensions right. So, you have a 3D, 3D you know representation of the real world which is unlike our previous two examples right.

So, what is happening here is that you have a x-axis and a y-axis right, which is giving you the domain of land in a region that is being captured in the image that we are looking at. What we are also seeing is height of trees, the height of this house, little house on our image and you know and there is a differential that we can see of heights between these different objects.

So, different different trees and between the house and other objects in the picture. So, when you look at the raster format or the raster understanding of it, what you see is a pixelated or a cell-like format created with rows and columns right. So, again you have these rows, you have columns and what you are doing is that you are using, you are giving a value to each cell to represent what is happening in the real world.

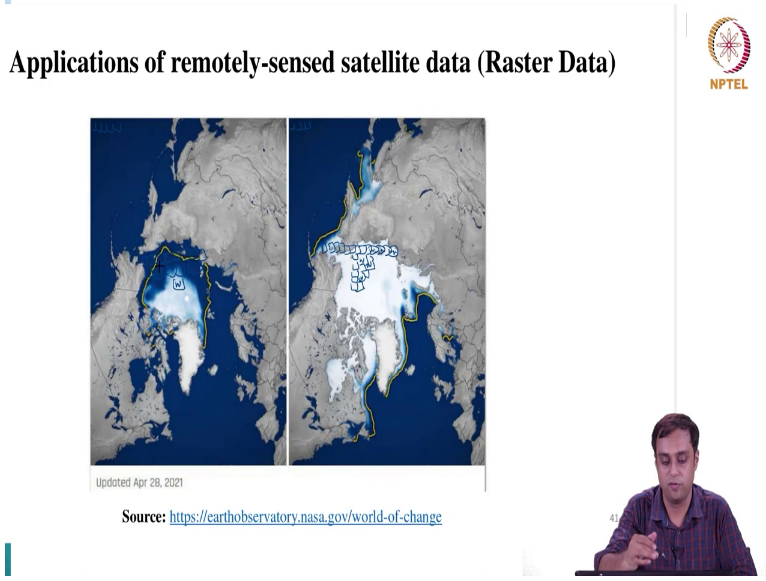
So, where wherever you have R, you are giving a value R that is where you see a river ok. These are cells which are representing the river in my data set right, in my real world data set ok and then you have you have these Ps which is a cluster of trees which are near the river and these S values which are a cluster of trees away from the river right and then you have a house which is one cell on the picture right.

So, that is the raster, you know, representation of the real world. When we come to the vector representation very interestingly the river is reduced to a line just like a road, in the previous example the river is reduced to a line. Once again you have an x and y axis, you have a coordinate system that you are working with and you have a gridded format that you create with discrete steps 1, 2, 3, 4, 5 both on x axis and y axis.

And, through this, through these gridded points you simply create a data structure, where you reduce the curvy looking sort of you know river into a line with 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 vertices, right. The house is reduced to a point and the two clusters of these wilderness areas which are cluster of trees are now sort of, you know, reduced to polygons with different shapes representing the clustering sort of arrangement in space.

And, then also representing the location in the sense some the cluster of trees P denoted by P are nearer to the river that is on the river banks and then S is away from the river ok.

(Refer Slide Time: 17:49)



So, having now understood in detail how a raster data format works versus a vector data format, we will now quickly go over some of the applications that we come across right as in raster as well as the vector format. The examples that we have seen previously will also apply to this particular classification and we look at applications of remotely sensed data set or spatially delineated data set in raster format as well as in, you know, the vector format.

So, what you see here is an image of the Arctic, more particularly the Arctic ice and you see that you know between two seasons that is a winter and the summer, you can see the difference in the size of the Arctic ice in April 2021 right. So, you can visualize how the ice structure, the ice mass shrinks and expands between these two seasons. And, this is done using this continuous format where each cell in the image is containing a digital value.

When that is happening what happens around you know in winter you have a certain value let us say W for all these pixels that that I am drawing on your on the Arctic ice, that is the way you know our raster data are organized right. When you overlay these pixels right, when you overlay these pixels for, you know, for the summer season what happens is that many of these locations turn from white to blue and; that means, the ice has melted in that region right.

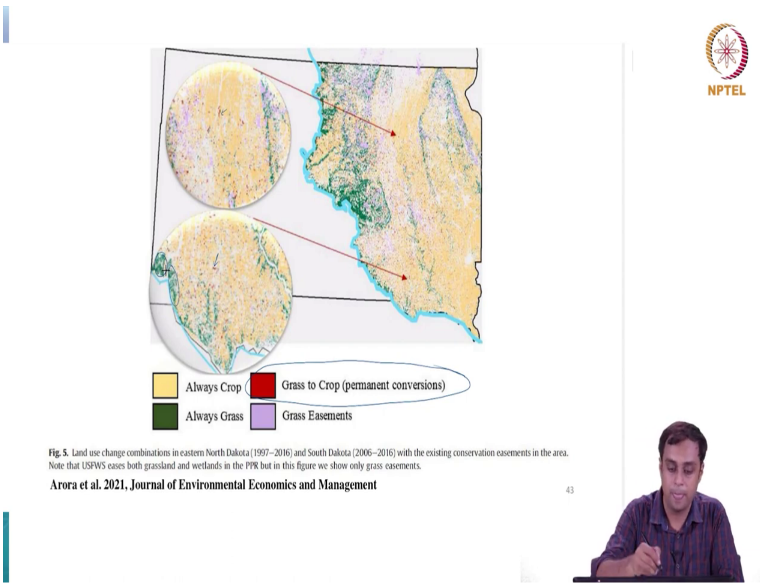
So, you are able to study the change in ice mass by looking at the raster representation of the data. Similarly, this is the Amazon; so, this is deforestation in the Amazon.

(Refer Slide Time: 19:48)

You are looking at images that are ranging from 2000, year 2000 to year 2005 and year 2012. Everywhere you see that you have this, you know, expansion of the brown color in the raster format on raster pixels you have basically replaced a green pixel which was the Amazonian forest to some kind of a development; it could be an agricultural land, it could be a road and so on and so forth.

So, what you see here is an intensification of, you know, deforestation through time in a region, in a particular region you know in Amazon in the Amazon rainforest.

(Refer Slide Time: 20:38)



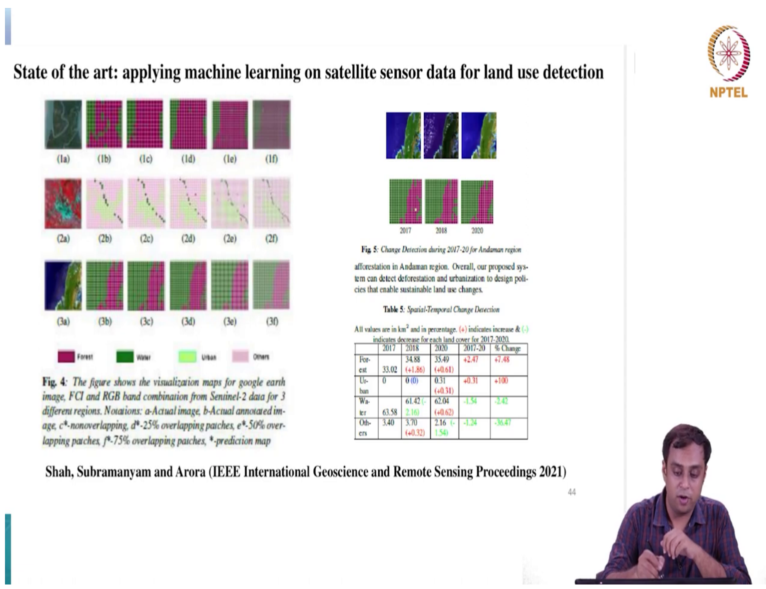
This is another image which is also using raster data, on this image you have you know categories like always crop, always grass. So, you have grasslands, you have crop lands, they are existing besides each other. And among these are very interesting dynamics are grassland to cropland conversion. So, if you are trying to understand decisions as to when or where exactly conversion decisions happen; this image tells you that the decision to convert happens as islands, as islands in large cropland tracts right.

So, you have large, you know, grassland tracts, you do not do not see too many reds there right. So, the reds appear over time on regions which are yellow or in between within those regions which are yellow; that means, that you are looking at conversions happening in a region where the land is already cropland you know in the near locality right. This can happen due to many reasons, you know you have markets which are more matured, you have more agricultural services.

For example, you can rent agricultural machinery like tractors, you have services like insurance provision, you have credit, you have banks etcetera right. So, if you have a very sort of a cropland sort of intense region, there is already an infrastructure, there are highways, there are transportation you know infrastructure, there are *mandis*.

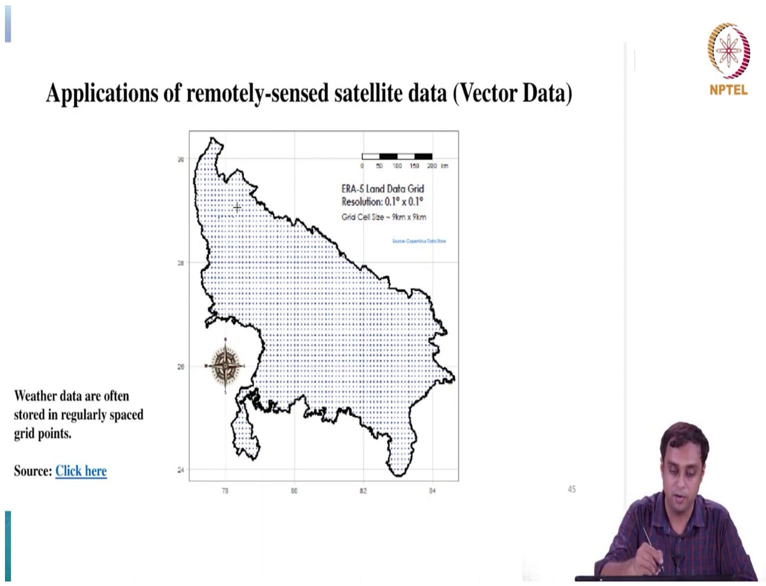
So, there is this infrastructure which sort of incentivizes, you know, let us say landowners or farmers to then convert their land towards cropland you know in regions which are already under you know highly cropped areas.

(Refer Slide Time: 22:18)



I think this is the last example. Now, this is where the, you know, the raster image is at. So, there is a lot of application of machine learning on raster data and in identifying land use changes. This is just an example where you know a study is trying to sort of you know identify different types of land uses in the Andaman region of India. You can look at it if you are; if you are interested.

(Refer Slide Time: 22:44)



So, finally, you know I want to sort of go over some of the applications of vector data. So, just like you know I had said earlier vector data are stored as points, lines or polygons. One of the most popular or prominent sort of applications or you know uses of vector data is in storing weather data. So, weather data are stored and weather stations on weather stations or in gridded format like you are seeing on the screen here. So, you have these points placed equidistant from each other along both x-axis and y-axis right.

This is the region of Uttar Pradesh and what you are looking at is that each point in space has a daily record of you know minimum temperature, maximum temperature, precipitation and so on and so forth. So, the data are stored at this point location, this is a storage mechanism right. It also allows us to sort of visualize how the data would look like only through what the value that is attributable to these points on the map in front of you.

(Refer Slide Time: 23:59)

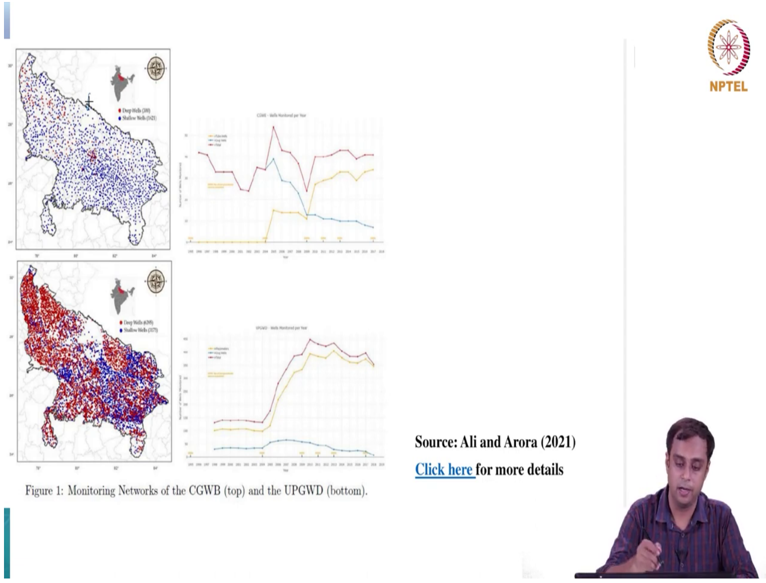
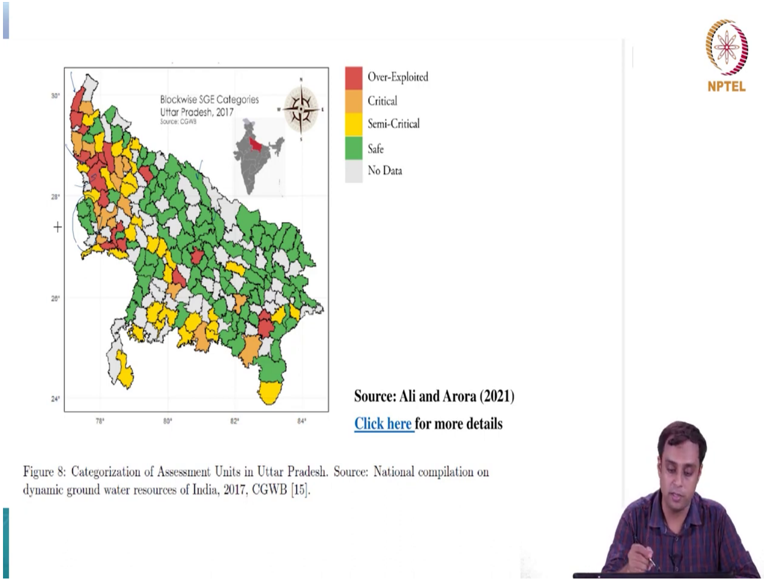


Figure 1: Monitoring Networks of the CGWB (top) and the UPGWD (bottom).

Similarly, you have a ground water monitoring network of the Central Ground Water Board and the Uttar Pradesh Ground Water Board here and each dot is nothing but a monitoring well you know that is owned by the ground water monitoring agency right. And, you can see how they are distributed in space, all these locations are used to monitor what is the level of ground water at any point of time that is an interest of the agency right.

So, the agency can and walk in, walk up to a one particular well, monitor what is the level and then come back later and figure out what is the level. So, if you see water level going down, you have a indication of depletion right. So, such data sets are also vector data right.

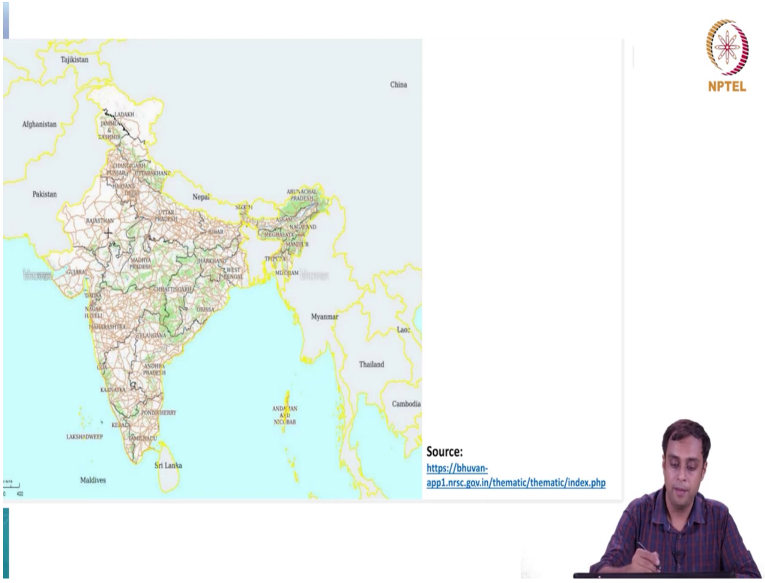
(Refer Slide Time: 24:44)



I talked about administrative boundaries. So, here is a block-level you know assessment of you know Uttar Pradesh again in terms of ground water levels and the greens here are regions which are safe in terms of you know groundwater depletion. The yellows are somewhat in worse situation, they are semi-critical, the oranges are you know are in trouble and the reds are over-exploited; so, you know that.

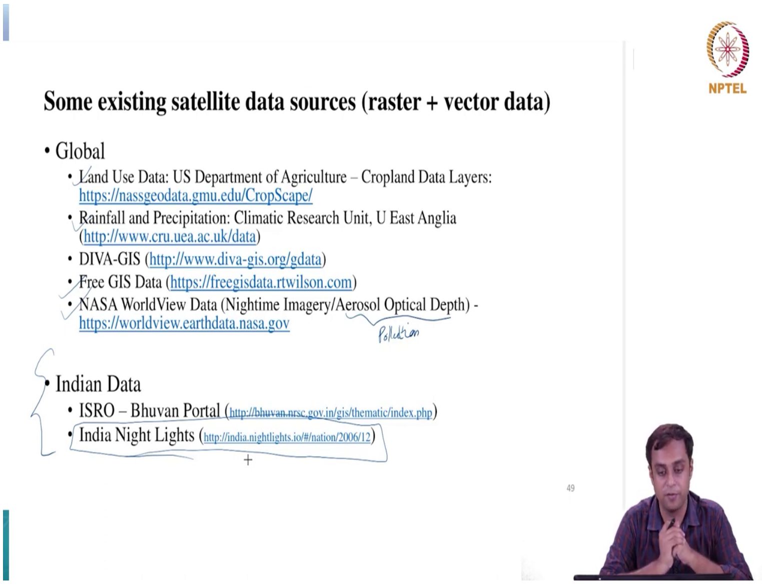
So, you see here on the western Uttar Pradesh side which is closer to the national capital region you have a lot more red that is over-exploited regions than others. That does not mean that they are all over exploited, you also have green regions around the NCR which suggests that there is a there is quite a bit variation in how you know ground water dynamics works and these data are recent data. They are 2017 data; so, they are not; they are not ancient data or something.

(Refer Slide Time: 25:40)



Here you see again the line, you know, format of vector data being used to visualize the national highway network of India. So, all the brown sort of lines that you see on this map are nothing, but vector representations, vector storage of you know spatially delineated road, national highway network in India, ok.

(Refer Slide Time: 26:07)



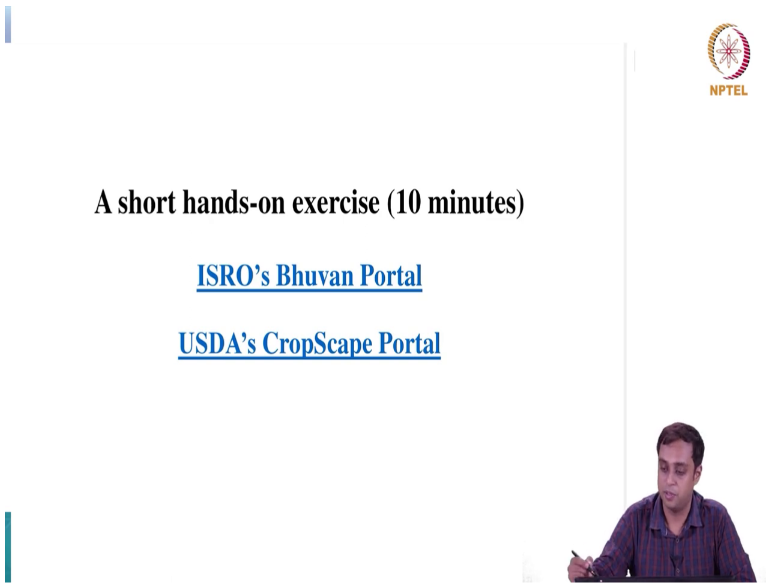
So, I want to sort of, you know, take you to some of the existing popular you know data sources of both raster and vector data. So, these data sets that I have been showing you know over time; you know they are all most of them are freely available online. I am providing you some of the sources here. As a next step what we will do is we will actually go over some of the data.

And, we will do some exercises to understand and you know how we can use this data, how we can bring this information to our own analysis; you know if you are researchers, you are working in policy, you work in consulting or you know you want to sort of explore some of those, some of those domains. Then you know these data sets you know the land use data set of the US Department of Agriculture, the rainfall and precipitation data set of the Climate Research Unit of University of East Anglia; a version of which we saw for Uttar Pradesh a while ago.

There are these free GIS data, there is you know the night-time imagery, aerosol optical depth which is a representation of air pollution. These are provisioned by NASA and its freely available at http://worldview.earthdata.nasa.gov . There is also Indian data which is, you know, made available by Indian agencies. The most interesting of those which is actually interesting to me is the Indian Space Research Organization's Bhuvan portal.

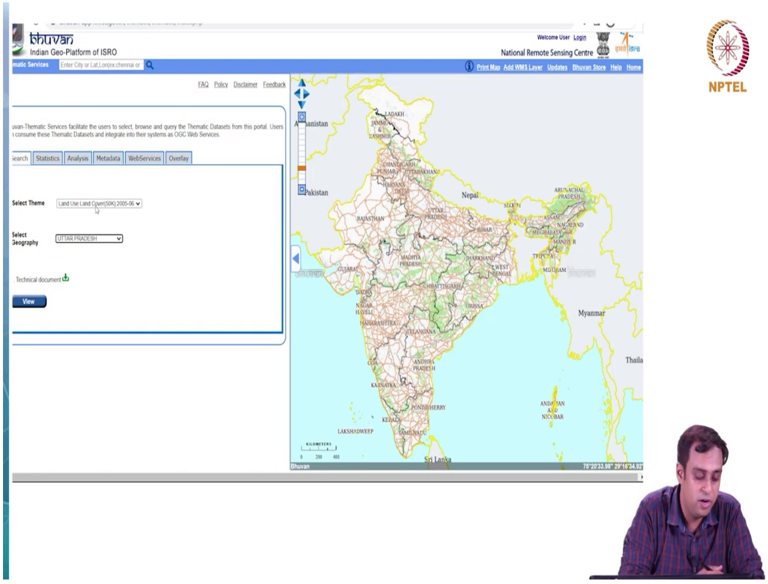
We will look at this data in some detail, it has many interesting features and then there is India nightlights data which we saw during the you know starting sort of you know beginning of this course, where I showed you data on nightlights and how they are changing over time in India. It is also a part of this data set is also sort of used to sort of delineate the economic development metrics for the economic survey of India ok.

(Refer Slide Time: 28:16)



So, as a next step we are going to do a short hands on exercise with two of these data sets; so, that you know we can see that you know we can actually work with these data set. They are not really abstract entities, they are, they are very much usable you know as available, ok.

(Refer Slide Time: 28:38)



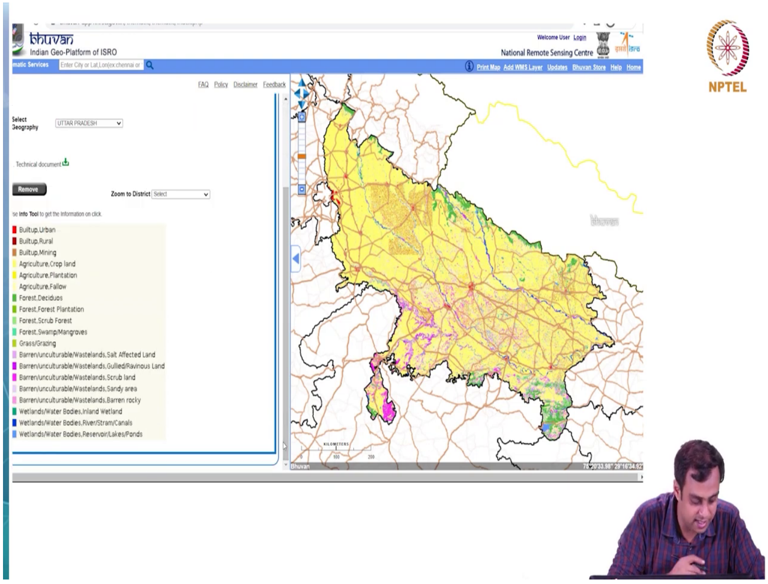
So, we are looking at on your screen we are looking at the Bhuvan land use land cover data set. The URL of which was pointed out, you know, in which I pointed out to you on the slide. So, here what you see on the right is a map of India, we see a layer which is overlaid here which is the national highways which are in brown you know lines which is a vector format.

And, then we have these state names within India and then outside India we have different countries which are also you know delineated here. So, when we start working with these data, the first thing that we need to do is we need to select a theme on which we want to analyze the data. So, the first theme you look at is land use land cover which is 2005-06, land use land cover 2011-12, land use land cover 2015-16.

So, you have these data layers for different different years available for direct visualization on the website of the Bhuvan data set, right. You have also other interesting land use types which is let us say land degradation, what is the state of land degradation, what is the urban land use status, wasteland, glacial lakes, flood hazard, so on and so forth.

So, you should, you know explore all these types at your disposal right. I am going to go with the first one which is land use stand cover 2005-06. Once I click this, I get a select geography option, when I do when I go ahead and try to select what I get is different states in India. Because, we have been looking at you know Uttar Pradesh some examples, I am going to go and look at Uttar Pradesh.

(Refer Slide Time: 30:28)



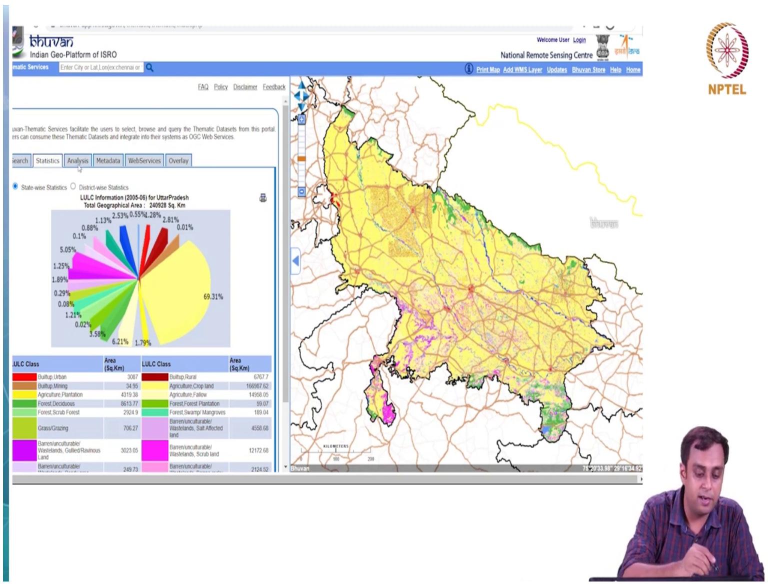
So, I select Uttar Pradesh as a geography of interest and then I can click on “View.” When I once I click on “View” on the right this website directly points to land use in Uttar Pradesh you know that I am looking at right. This is for the year 2005-06, when we say 2005-06 it is the cropping calendar of India that starts right at the time the monsoon comes; that is May of every year right, May or June of every year and then goes [FL] next year that is 2006, just before the monsoon let us say till April of 2006.

So, May 2005 to April 2006 is what is called as 2005-06 ok, and what you see right away when you click view, you see legend of different colors, different digital values on this map and what they represent. So, first of all this map of India is really a raster format right. So, it is a continuous data structure, right, if it is a pixelated data structure and for every pixel on my image I have a value which is represented by a unique color right.

If you have two pixels which are side by side, have the same color that basically means you have a similar type of land use going on in contiguity in that region right. If you have a jump in colors between two pixels you have a; you have a some kind of a boundary or a discrete jump in the land use type that is urban to rural that or agricultural. So, if you have a urban or agricultural frontier then you have these jumps between red and yellows, right given this given this legend.

So, yellows are obviously; so, you have reds and browns which are urbans built-up areas, the yellows are agricultural lands, the greens are forests, you also have grass. There are other kinds, you know, there is barren land, there are wetlands, there are water bodies and so on and forth ok. So, this is the legend.

(Refer Slide Time: 32:36)

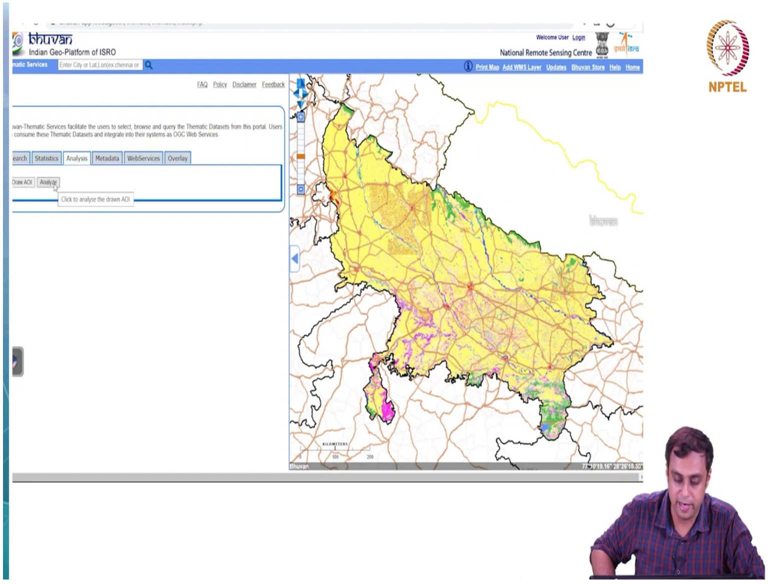


If we click on statistics, we will get a nice pie-chart of what is the break-up of land use type in 2005 and 06 in Uttar Pradesh. These types of pie charts are what you are used to right, if you are given data of what is the percentage of agricultural land in UP, urban land in UP, wasteland in UP, water bodies in UP, forest area and UP and so on and so forth; you can create a pie-chart on excel right that is not that is I am sure all of you are aware of it and its quite straightforward.

What is interesting is that we are creating this pie chart from an image in this particular case right. So, now, we are able to visualize that the data that we are used to working with as data scientists or as applied statisticians as econometricians on a Comma Separated Value (CSV) format, CSV format in an Excel sheet can indeed be basically mapped to an image.

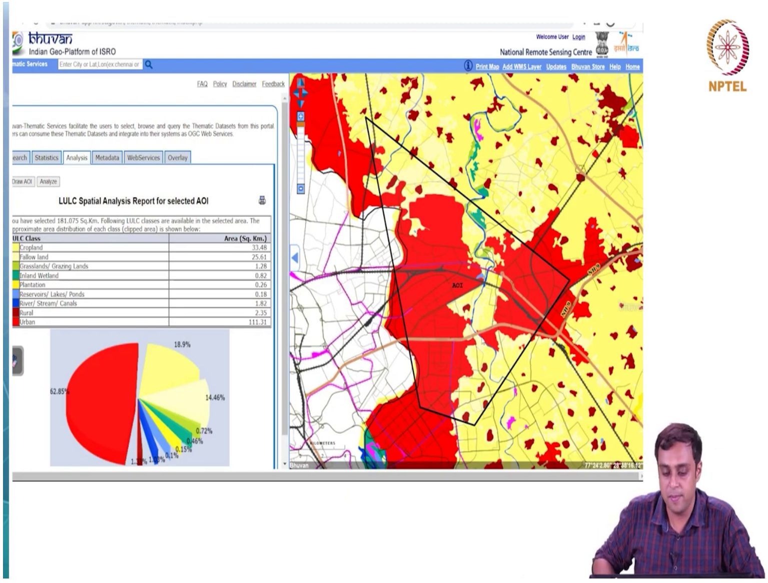
And, the data that we see on an image can be then mapped to the CSV format or other similar formats right; that is interesting. What you also see is area in square kilometers, we can go one step further.

(Refer Slide Time: 33:47)



We can click on analysis and here what it says is you can draw a you know AOI and analyze. So, I am going to click draw AOI. So, we click on draw AOI, you can see my pointer AOI means Area Of Interest. We click here and then we go to the map and I go to the Western UP region which is adjoining the NCR. And, I will draw a small area of interest which is let us say near you know NOIDA a on the east of Delhi right.

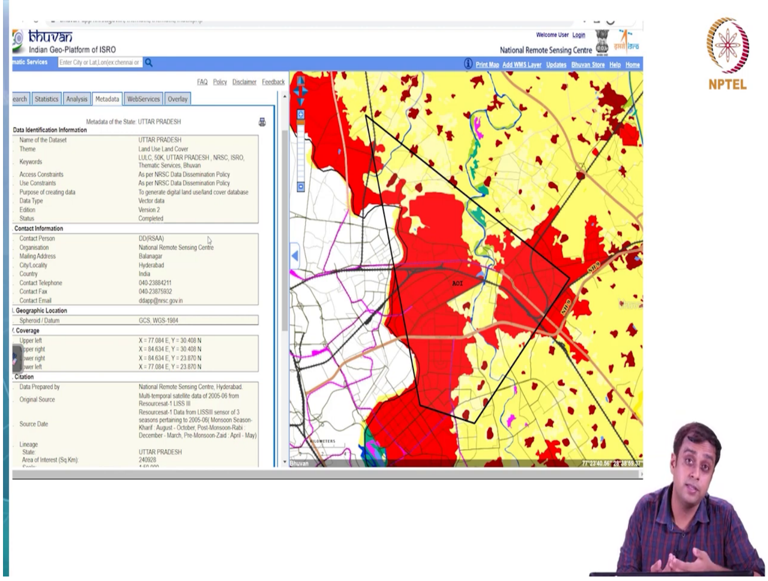
(Refer Slide Time: 34:27)



So, I am going to click "Analyze", come back and click "Analyze" and what you see here is that my AOI that I had drawn on the image now appears on this map and I have statistics for the area of interest that I have drawn on my image right. So, you can look at an entire state, but you can actually then go to a local region of interest which as a researcher, as someone who is just curious wants to know.

And, we have now come to a region which is 62.85 percent urban land right, we also have quite a bit of crop land here which is let us say 20 percent which is 18.9 percent. We have fallow land, we have grasslands, we have lakes, ponds, river, you know we have inland, wetlands. So, there is a; there is a very diverse kind of a land use even in that local you know in this local area of interest that we are studying here. As a next step we can move to what is called as the metadata.

(Refer Slide Time: 35:28)

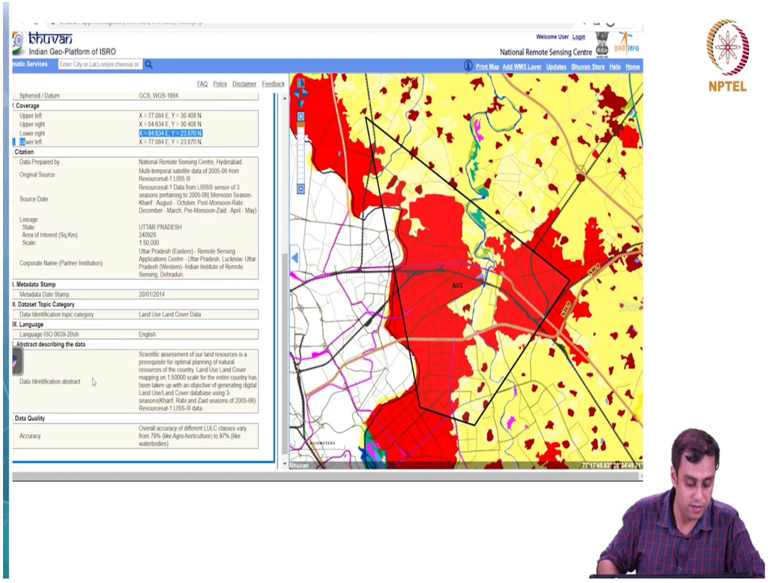


So, the metadata are the ones which contain all necessary information that you should know about this data including its data generating process, what is the level of accuracy. If you are using these data you must cite these data, how should you cite these data, if you have an issue with this data whom should you contact right, if you have queries about these data whom should you contact right. So, let us look at it.

So, there is data identification information, name of the data set, what is the theme, the keywords, access constraints right, use constraints and so on and so forth. There is contact information just like I said, there is geographic location right. So, we have a GCS - a Geographic Coordinate System; remember we have studied GCS in our in our lecture, in lecture 2 right. We have a coverage which tells us what are the extreme x y coordinates on upper left on you know upper right, lower right and lower left.

So, in this AOI upper left, upper right you know lower right and then lower left. So, it is giving me the x y coordinates that that encompass the area of interest right, then we have citation. So, if you are using these data this is how we cite these data right.

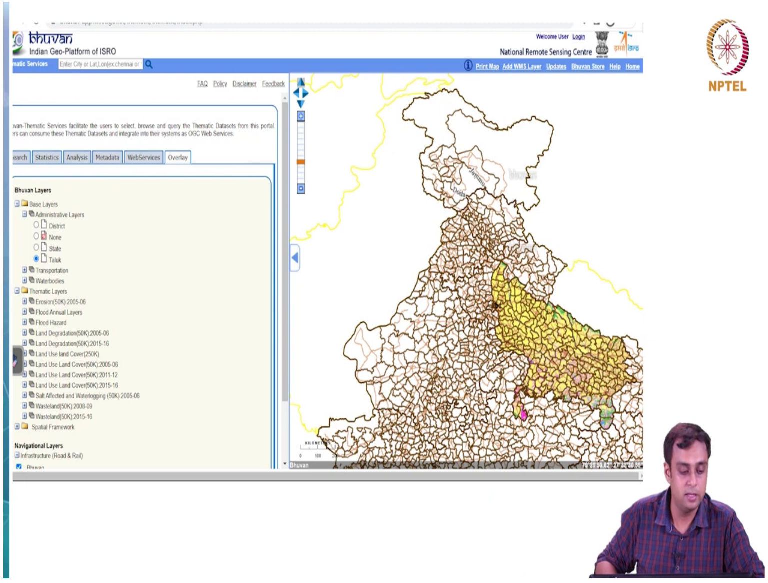
(Refer Slide Time: 36:49)



So, we must be aware of this. One very interesting you know data quality parameter that metadata provides that is something that we should document whenever we are using these data is on accuracy right. It says overall accuracy of different LULC classes vary from 79 percent like agro, horticulture. So, whenever we see a agricultural classification it is true only 79 percent of the of 100 times right.

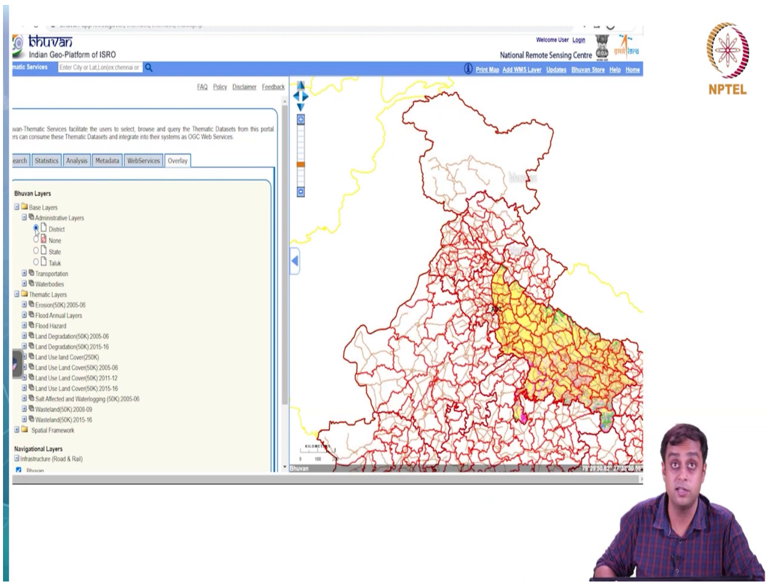
So, 79 of 100 times we are sure, we are confident that we have which we have that class, but for other for you know remaining 21 times, it could be some other class. So, there is some error in these data right. So, we should we should be very careful about these and we will study how to articulate these errors you know analytically. And, then we have a 97 percent you know accuracy for water bodies. So, we are doing a very good job with articulating where the water bodies are.

(Refer Slide Time: 37:45)



Finally, I want to look at this overlay function on this on this website. Here what you can see very interestingly is that you can overlay you know administrative units layers like taluk. So, you can study you know the Uttar Pradesh region at a much you know finer scale that is taluks.

(Refer Slide Time: 38:06)



You can move to districts. So, you can see districts are larger than taluk. So, taluks are more finer resolution you know administrative units right. So, that is about it, I would like for you to look at these data, work with these data, you know play around a little bit. We have we will do a little exercise; you know in the next lecture.

I want you to sort of go and look at the Guwahati region and look at the Guwahati metropolitan area, draw area of interest and analyze land use types, look at the metadata and then maybe overlay a few features right. And, we will sort of do that again in the next lecture along with CropScape, that is another publicly available, freely downloadable data set we will see the functionalities for it.

Thank you.