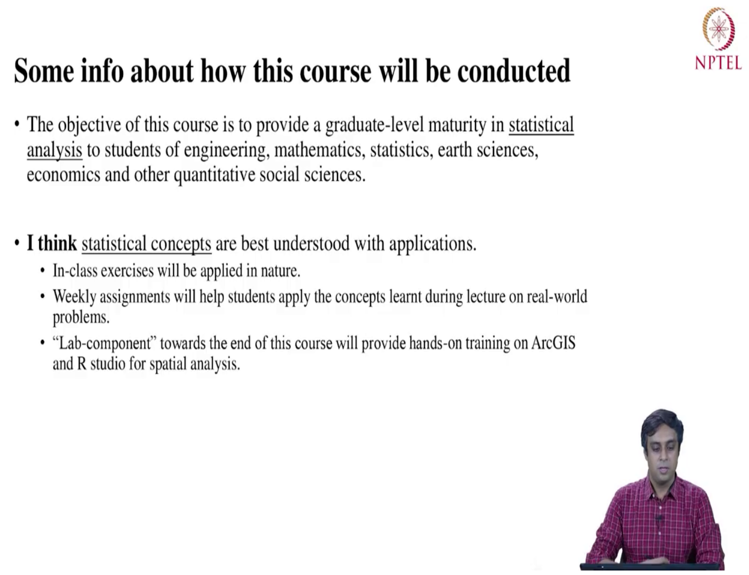
Spatial Statistics and Spatial Econometrics

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

Lecture - 01
Introduction to Spatial Data Analysis

Hello everyone. My name is Gaurav Arora, I am a faculty of Economics at IIIT Delhi and let us get started with the first lecture of Spatial Statistics and Spatial Econometrics.

(Refer Slide Time: 00:34)



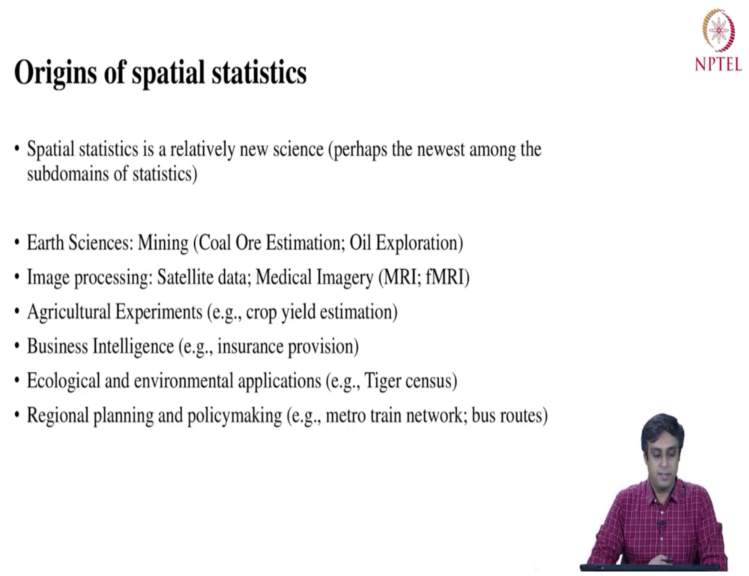
Some info about how this course will be conducted

- The objective of this course is to provide a graduate-level maturity in statistical analysis to students of engineering, mathematics, statistics, earth sciences, economics and other quantitative social sciences.
- **I think** statistical concepts are best understood with applications.
  - In-class exercises will be applied in nature.
  - Weekly assignments will help students apply the concepts learnt during lecture on real-world problems.
  - "Lab-component" towards the end of this course will provide hands-on training on ArcGIS and R studio for spatial analysis.

So, we will begin with some information about how this course will be conducted. So, the primary objective of this course is to provide a graduate level maturity in statistical analysis for students of engineering, mathematics, statistics, earth sciences, economics and other quantitative social sciences. So, at the outset we can see that this course will have wide applications. So, you know a wide range of students coming from different domains can take advantage from this course.

I believe that statistical concepts are best understood with applications. So, in this course we will strive to have you know in class exercises which will be applied in nature. We will have weekly assignments that will help students apply the concepts that are going to be learnt during lecture hours on real world problems. And then there will be a lab component towards the end of this course which will provide hands-on training on GIS software using ArcGIS as well as some packages in R that have similar or even more advanced, you know, applications.

(Refer Slide Time: 01:49)



Origins of spatial statistics

- Spatial statistics is a relatively new science (perhaps the newest among the subdomains of statistics)

- Earth Sciences: Mining (Coal Ore Estimation; Oil Exploration)
- Image processing: Satellite data; Medical Imagery (MRI; fMRI)
- Agricultural Experiments (e.g., crop yield estimation)
- Business Intelligence (e.g., insurance provision)
- Ecological and environmental applications (e.g., Tiger census)
- Regional planning and policymaking (e.g., metro train network; bus routes)

So, about the origins of spatial statistics. Spatial statistics is a relatively new science. Perhaps the newest among the subdomains of statistics. There are several other subdomains of statistics for example, there is biostatistics among these subdomains spatial statistics seems to be the newest ones.

The early applications of spatial statistics came from earth sciences specifically in mining where for coal ore estimation that is the quality of coal available or any ore available inside the earth is fundamentally not visible to the statistician, the miner, the analyst that is out there to dig that or extract that coal.

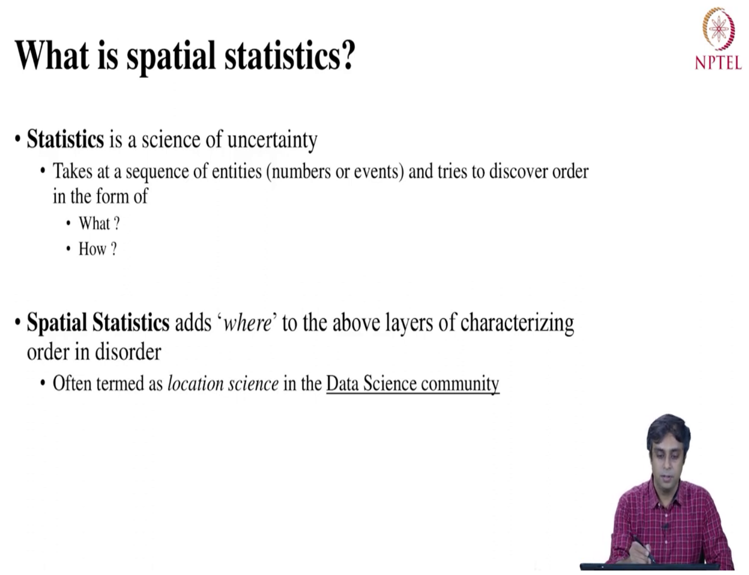
So, what we can achieve at best is that we can dig some holes at some locations on ground and we have to predict what is the pattern of quality underneath the ground from what we can observe from few points on top of the earth where we have dug. Similarly, you have problems of oil exploration.

So, oil search campaigns you know done by petroleum industry have similar issues you know where should they send their next expedition right, these are costly expeditions. So, that is where statistics comes in. It tries to predict what would be the best optimal pathway in terms of extracting coal from ground or searching for oil.

More modern applications are of image processing therein we will look at satellite image data that is one of the most popular you know applications of spatial statistics there is medical imagery MRI, FMRI that is where that is also where spatial statistics is finding its use today. In crop yield estimation for agricultural experiments, for insurance provision by business intelligence consulting firms, for conducting tiger census you know by ecological and environmental scientist is also finding its use of spatial statistics.

Finally, for regional planning for example, planning bus routes in a city, planning a metro rain train, sorry metro train network, you know where to put the next route you know what should be the best you know locations where you know most residents of a city can find connectivity will find applications of spatial statistics.

(Refer Slide Time: 04:12)



What is spatial statistics?

- **Statistics** is a science of uncertainty
  - Takes at a sequence of entities (numbers or events) and tries to discover order in the form of
    - What ?
    - How ?
- **Spatial Statistics** adds '*where*' to the above layers of characterizing order in disorder
  - Often termed as *location science* in the Data Science community

So, with that you know let us define formally what is spatial statistics. So, first of all before we get to what is spatial statistics we should understand what is statistics. Well statistics is a science of uncertainty. So, statistics posits the observations that we see around us in the world as fundamentally random in nature right. But you know in this fundamentally stochastic or random real world around us there is some order and statistics as a discipline as a science strives to determine or identify order in disorder, right.

So, it takes a sequence of entities or sequence of events for example, climate change right if you want to understand the pattern of drought. So, fundamentally statistics posits "incidence of drought" as a random event. What do we see is some observations of drought in the past right. So, let us say a drought happens in the Indo-Gangetic plains once in every 3 years. Sometimes it happens within 2 years, sometimes it happens after 5 in 5 intervals right.

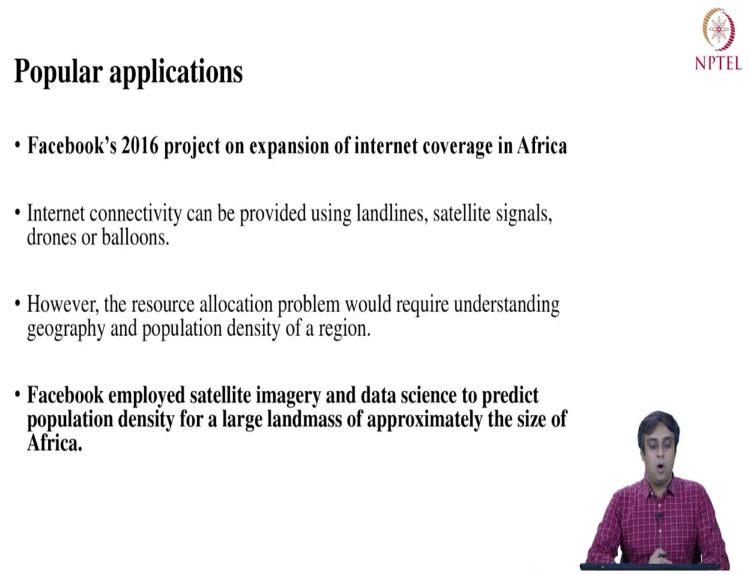
So, statistics would take the sequence of events that have been observed in the pasts and then try to discover some kind of order in the sense of what is the average frequency the you know median frequency or similar type of moments between two different droughts and you know what kind of events precede this drought ok. So, how can we predict the next drought? That is the basic, you know, basic line of inquiry.

Spatial statistics adds a new dimension of *where* to this science of what and how in determining order in disorder. So, it would then ask, not only when will drought happen and how it will happen, how strong it will be in terms of its in intensity, it will also ask *where* will it happen exactly in space. So, we bring in a new dimension of space if you want a little bit more intuition you can think of the dimension of time you know. So, time itself is a new dimension as far as you know time series statistics is concerned.

In time series you have a time origin t_0 then you have t_1 which comes after t_0, then t_2 that comes after t_1 and so on and so forth, it keeps going till infinity right. In space you also have a point of origin, but then you can actually move in any direction from that point of origin as far you know that we can imagine you know the four directions are North, South, West, East that is something very natural that comes to us.

So, if we start to exploit these directions in space in doing statistics that is what you know that is the toolkit that spatial statistics will bring to the table ok.

(Refer Slide Time: 07:12)



Popular applications

- **Facebook's 2016 project on expansion of internet coverage in Africa**
- Internet connectivity can be provided using landlines, satellite signals, drones or balloons.
- However, the resource allocation problem would require understanding geography and population density of a region.
- **Facebook employed satellite imagery and data science to predict population density for a large landmass of approximately the size of Africa.**

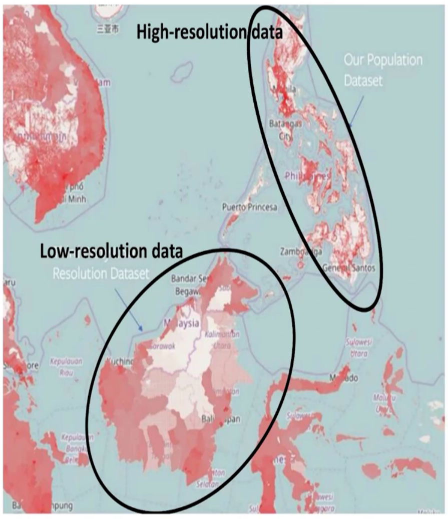
So, let us begin with some popular applications alright. So, you know, because we have to sort of combine concepts with the applications in that spirit, you know, let us see who cares about spatial statistics. So, some of the most recent popular applications of spatial statistics have been you know it has been a project by Facebook on expansion of internet coverage in Africa.

Now internet connectivity can be provided using different modes, you know it can be provided using landlines, using satellite signals, using drones or even hot air balloons. However, which mode is best suited for which region depends on the demand for internet which will come from the population density of that particular region.

So, if we are going to remote regions let us say of Africa or even in India where we do not really have a sense of population density then what do we do? Well, then we can use satellites and data that comes from satellites to predict population density you know of a region which may not be otherwise known and hence we cannot, you know, get internet to that particular population or that community at the best you know in the most optimal manner.

So, Facebook employed satellite imagery and data science to predict population density for a large landmass of approximately the size of Africa. So, it was a large-scale application done using satellite imagery.

(Refer Slide Time: 08:41)



Source here

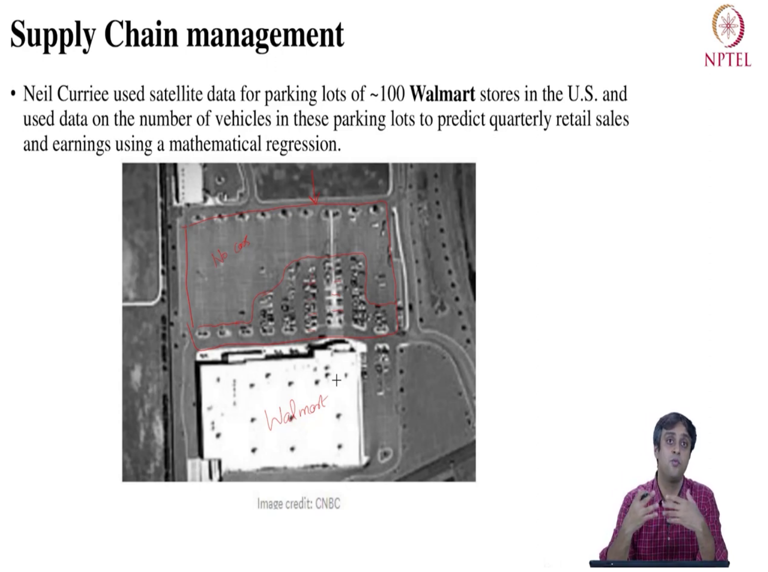
And here is, you know, a particular data product that has come from this project of Facebook. Well what you see is that on the north, east side of this you know of this picture or this map what you see is a high resolution data that you know Facebook predicted using satellite imagery.

And on the south you have a low resolution data set that comes from administrative surveys. What you see is that in the bottom picture you have a sense of population density in the sense the lighter regions are low population density areas the darker regions, the bright red regions are high population density areas.

But in the bottom picture you know a particular district can have a uniform level of, you know, population density, right. So, if you have a white block or a white cell it is saying that the population density is of certain level for this entire block right. Whereas, if you go to this north map, this map in the north, you can have difference in the level of population density in that in a block of similar size ok.

So, you can have a you can have lower density areas adjoining higher population density areas. So, you have higher resolution, higher resolution understanding of the region by itself. So, this is powerful in the sense that you know it gives you a more sort of finer understanding of what is happening on the ground, so that you can then allocate resources efficiently.

(Refer Slide Time: 10:28)



Supply Chain management

- Neil Curriee used satellite data for parking lots of ~100 **Walmart** stores in the U.S. and used data on the number of vehicles in these parking lots to predict quarterly retail sales and earnings using a mathematical regression.

Image credit: CNBC

Let us look at another example. So, in the in another popular example of supply chain management which is a more of a business intelligence domain example. Satellite data were used to estimate the number of vehicles parked or density of parking area utilized for 100 Walmart stores across the United States. And then these were then correlated with those with the revenue streams and the sales of those stores.

So, what you see in this picture is that you can you are looking at a satellite imagery which has a parking lot, it has a parking lot and you know some portion of that lot has cars in it, right. So, you can first of all see through naked eye that you know some of some portion has cars some portion has no cars right.

The other thing that you can see is that you can imagine that one can use software to actually count the number of vehicles that are parked in this parking lot. So, you can tell at a given point of time how many vehicles are being are parked in this parking lot which gives you a signal of how many customers, how many consumers might be visiting that Walmart store, this Walmart store at that given point of time. And that should then give you a signal of what should be the level of sales or level of you know earnings that this store may be, you know, experiencing.

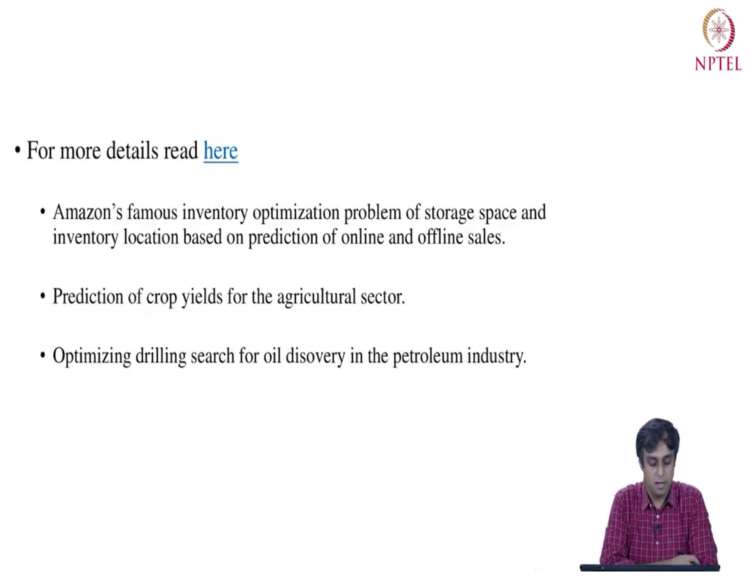
So, satellite data are can be used in such a way for business intelligence you know purposes.

(Refer Slide Time: 12:11)

Economic Survey of India, 2021-2022

Tracking Development through Satellite Imagery and Cartography

(Refer Slide Time: 12:15)



- For more details read here
  - Amazon’s famous inventory optimization problem of storage space and inventory location based on prediction of online and offline sales.
  - Prediction of crop yields for the agricultural sector.
  - Optimizing drilling search for oil disovery in the petroleum industry.

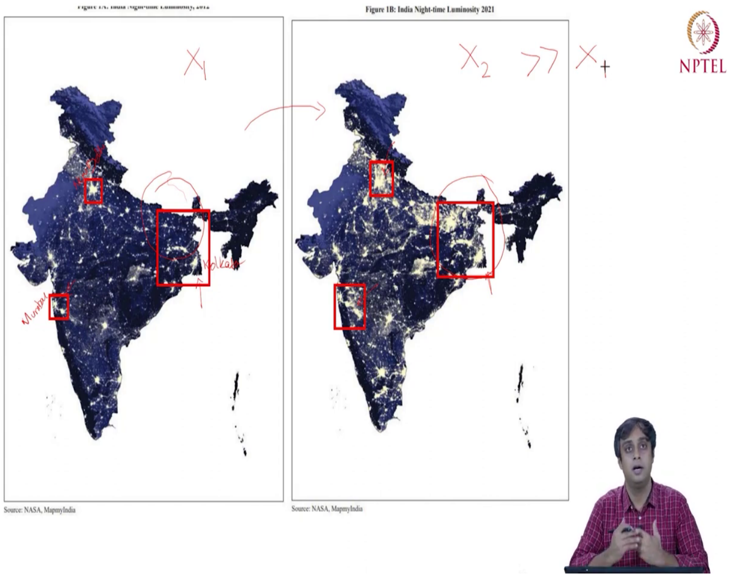
There are some other, you know, some other popular examples and I have a hyperlink here. So, if students are, you know, interested if you are interested please click on the hyperlink go and read there are more examples for example, there is Amazon's famous inventory optimization problem that you will find very interesting to read.

Then there is prediction of crop yields for agricultural sector. This has become a very famous very popular tool in the west which has larger farms for countries like India or low income countries you know prediction of crop yields is more challenging because we have smaller farms.

So, this is a problem you know that is being explored by the research community at the moment. And then finally, you know as we have also remarked earlier there is optimization of drilling search for oil discovery that is done in the petroleum industry. A recent you know use of satellite imagery has been found in the Economic Survey of India 2021-2022.

This is probably the first time a government entity is you know is evaluating you know development or tracking development as they say using satellite imagery and maps. So, let us see what we find in the Economic Survey of India 2021-2022.

(Refer Slide Time: 13:24)



So, what you see here is night time luminosity on the map of India across space in 2012 and on the left and 2021 on the right. So, on the left you know first of all wherever you see a brighter light that basically means that there is, let us say, you know, there is more incidence of lighting you know, it could be city lights, it could be residential lights, it could be street lights; whatever that is all of that is aggregating and as and if we click a picture from space you know its looks like how it tells us how light is you know spread at night time in these cities.

So, you see New Delhi of course, around New Delhi you have a; you have a you know, very large mass of light in 2012. So, you have around Kolkata and similarly around Mumbai right.

So, these are metro cities of course, you will also see this in around Bangalore, around Chennai. I am only talking about these, right.

So, first of all you know the spatial information is rich enough to tell you the density of the difference the variation and density of night lights in across different cities and different towns in India if you come from a smaller town maybe you can locate that in this map approximately and tell how much night light was there in 2012 at that time.

On the right hand side when we move to 2021, what is very clear is that the night time luminosity has on average increased in India, right. So, if I were to look at a non-spatial statistic I will have a night time luminosity index for 2012 and a night time luminosity index for 2021 and of course, 2021 index will be much higher than maybe even double than 2012 measure, right. So, that is one type of statistical analysis or statistical understanding of you know data-driven understanding of night time luminosity in India.

But you can do better. You can tell that the increase in night time luminosity has happened around the already existing cores in 2012.

So, luminosity has increased, but it has increased mainly around you know New Delhi, Mumbai and Kolkata right what has also happened is that the area that looks quite dark which is the Bihar, Jharkhand, West Bengal and let us say Eastern UP, Uttar Pradesh that is Eastern Uttar Pradesh it looks pretty dark around the 2012 period seems quite bright. So, this is where the increase, the net increase is going to be coming from a lot right.

So, spatial statistics is powerful in the sense of not only telling us what has happened, how much it has happened, but it is also telling us where has it happened really, ok. So, if I were to tell you ok the night light luminosity was x_1 in 2012 it was x_2 in 2021 and x_2 was quite large as compared to x_1.

But then where has this increase come from well it has come from the core of already existing you know high luminosity density around the metro cities metropolitan cities of India, but it is also coming from the heartlands the not so developed regions that are that is Bihar, West Bengal, Jharkhand and Eastern Uttar Pradesh right. So, this is the interesting dimension that spatial analysis brings to the table.

(Refer Slide Time: 17:11)

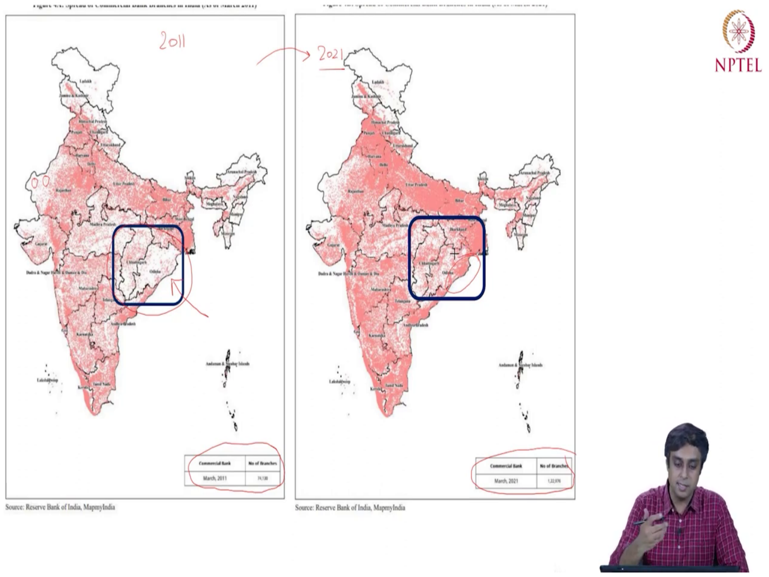


Figure 4A: Spread of Commercial Bank Branches in India (As of March 2011)

Commercial Bank	No of Branches
March, 2011	74,130

Source: Reserve Bank of India, MapmyIndia

Commercial Bank	No of Branches
March, 2021	1,22,976

Source: Reserve Bank of India, MapmyIndia

The next, you know, graph or the map that that Economic Survey of India posits is the spread of commercial bank branches in India. So, you can first, you can add the outset you can see the kind of you know applications that we are seeing as the first two examples in economic survey of India are very different; one is the night light luminosity the other is bank branches.

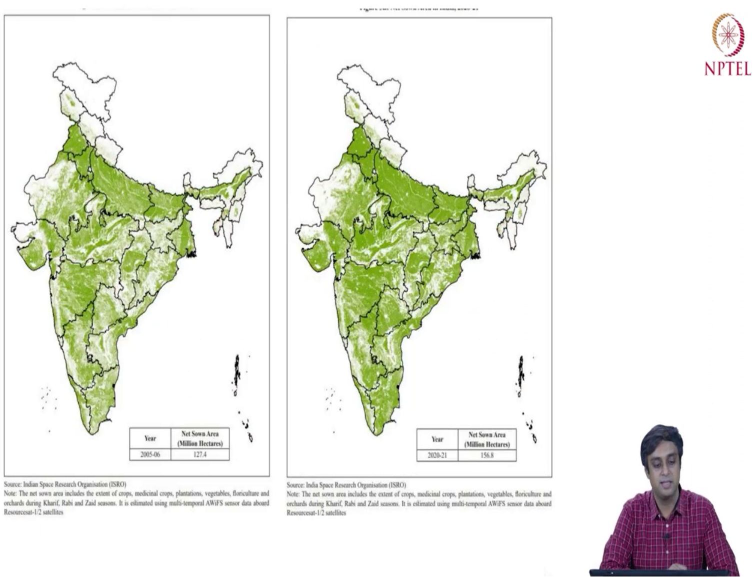
In this graph in the left graph we have the 2011, you know, spread of commercial bank branches every dot by itself is a bank branch. Every dot on this map is a bank branch. So, if you see higher density of dots what you are basically saying is there is high density of bank branches in that area. So, there is higher activity of economic you know transactions, you know the credit industry might be more, you know, dynamic in that region, right.

So, now you know on from 2011 then we have a commercial bank branches spread in 2021. At the outset what we are seeing is well there are definitely more bank branches in India in that from 2011 to 2021 right and the number of bank branches which is also coming from the economic survey has gone up from 74130 to 122976.

So, there is indeed, you know, a pretty high increase order 70 percent increase if, you know, if not double you know about 80 percent increase may be in, you know, in the increase of bank branches that is one way to understand the world around us right. So, how and how much by how much and what. So, what has happened? Well the number of bank branches has gone up by how much, well by about 80 70 to 80 percent right.

But where has it come from well, but this quite clearly coming from the region in Orissa and Chhattisgarh where you do not see you have this white map a whitish map which basically says there is less density of dots in this region and which goes up quite drastically at least in the coastal regions of Orissa right.

(Refer Slide Time: 19:33)



Year	Net Sown Area (Million Hectares)
2005-06	127.4

Source: Indian Space Research Organisation (ISRO)
Note: The net sown area includes the extent of crops, medicinal crops, plantations, vegetables, floriculture and orchards during Kharif, Rabi and Zaid seasons. It is estimated using multi-temporal AWiFS sensor data aboard Resourcesat-1/2 satellites

Year	Net Sown Area (Million Hectares)
2020-21	156.8

Source: India Space Research Organisation (ISRO)
Note: The net sown area includes the extent of crops, medicinal crops, plantations, vegetables, floriculture and orchards during Kharif, Rabi and Zaid seasons. It is estimated using multi-temporal AWiFS sensor data aboard Resourcesat-1/2 satellites

So, this is another very good example then we have another example on net zone area in India here you can see again you can have a statistics statistic of what has happened net zone area has gone up by how much you can calculate, but where has it gone up in the most? Well, you can see here that Jharkhand seems to have a quite drastic increase in net zone area relative to other states right.

You also see some of that in the Southern states that I have marked here. So, this is the type of extra information that spatial analysis is bringing to the table.

(Refer Slide Time: 20:08)

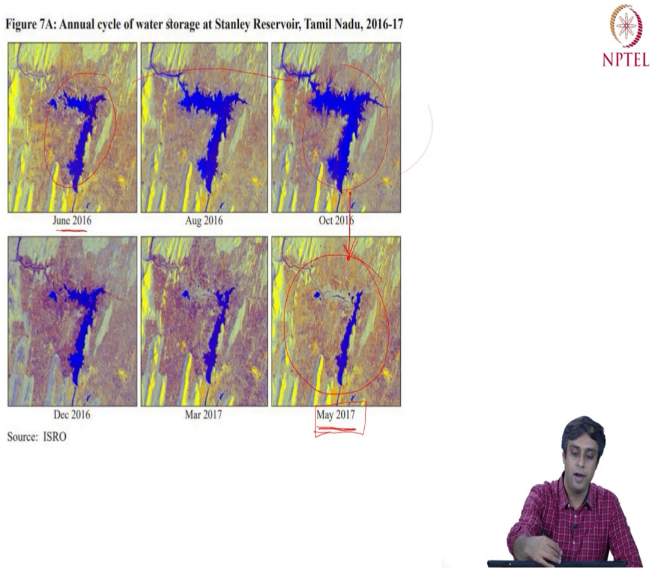


Figure 7A: Annual cycle of water storage at Stanley Reservoir, Tamil Nadu, 2016-17

Source: ISRO

Of course if net zone area is going up another there is also a downside to it that you can see shrinking of water resources. So, here you know we have a annual cycle of water storage at Stanley Reservoir in Tamil Nadu and we are looking at 2016 to 17. So, June of 2016 going up to May of 2017. So, June of 2016 is just you know right about at the onset of Monsoons.

So, we are looking at the size of the reservoir it goes up as the Monsoons you know play their role you know the rains happen and in October, you have a larger size of water you know contain capacity in this reservoir water storage in this reservoir. And then just before the onset of Monsoons you know in May, you see that the reservoir actually shrinks and the shrinkage from October 2016 to May 2017 would have happened due to let us say irrigation other types of demands of water, right.

So, this is another application of spatial data or data where we can find information which is delineated by space right which is let us say spatially delineated we can find applications in natural resource development right. So, if there is suppose if we are supposed to manage water resources probably we should look between post monsoon and right before monsoon period and try and conserve water the most because we see a lot of shrinkage during this period right.

(Refer Slide Time: 21:40)

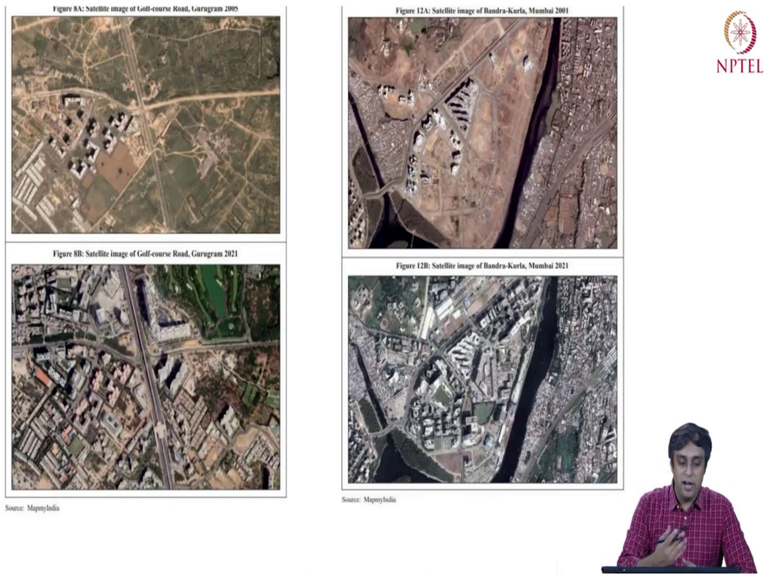


Figure 8A: Satellite image of Golf-course Road, Gurugram 2005

Figure 8B: Satellite image of Golf-course Road, Gurugram 2021

Source: MapmyIndia

Figure 12A: Satellite image of Bandra-Kurla, Mumbai 2001

Figure 12B: Satellite image of Bandra-Kurla, Mumbai 2021

Source: MapmyIndia

Similarly you have these urban development you know pictures on the left you have a Golf Course Road in Gurgaon in 2005 to 2021 you see the extent of development that has come about in 16 years period of time you know urban development of course, it has urbanized, but it has not urbanized everywhere if you look at the south west region that seems to have driven most of the urbanization whereas, the north east region seems to remain a lot of it remains seems to remain green right.

So, this is urban planning I mean this is supposed to be local you know local administration opening up land or allowing development or not or it could even be driven by the geographical features of the area right. There is a similar understanding which students can spend 30 seconds by looking at development in the Bandra Kurla area in Mumbai from 2001 to 2021.

And see where and try and identify where the where component of urban development in the sense you know has it happened you know where has it really happened? What are the areas it has not happened at? Does it make any impact on you know on the on the water resource that we see there you know about it does what kind of impact would it make on the greenery the green you know coverage of the area and so on and so forth ok. So, you can definitely try and identify these things on this picture on the right for Bandra Kurla Mumbai ok.

So welcome back you know we will look at couple more examples you know in this lecture.

(Refer Slide Time: 23:33)

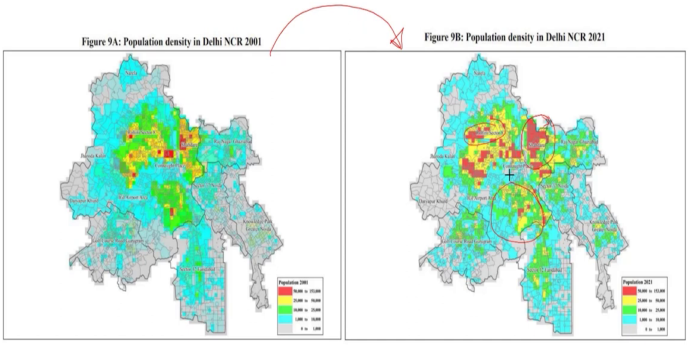
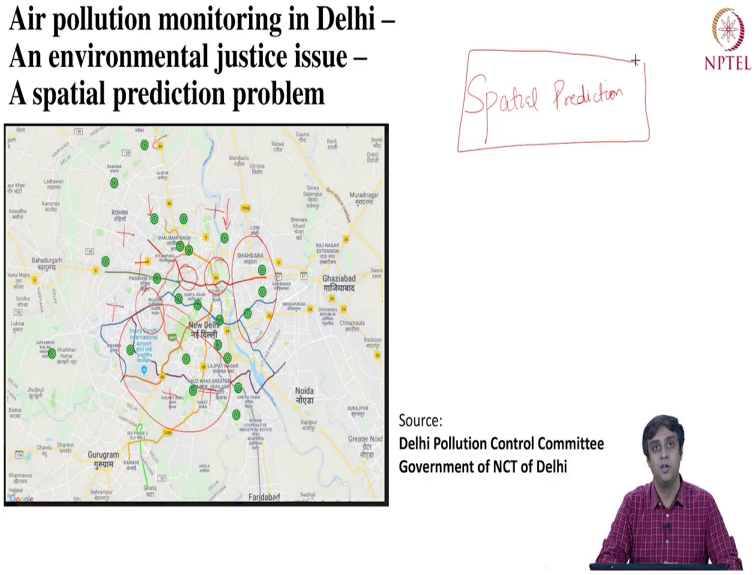


Image Source: MapmyIndia

So, the next example is you know the evolution of population density across Delhi. So, what you see is from 2001 to 2021 the first thing you can see is we have more red blocks in the 2021 map for the NCR region that is National Capital Region in Delhi then we had in you know 2001. So, we can say at the outset the population density of Delhi NCR has clearly increased from 2001 to 2021, but then we can go one step further like we have seen in previous examples that we can identify what are the regions where this has really happened.

So, clearly we have the region of Shahdara which has seen a large increase in or the or the dominant increase in population density among different regions of Delhi we have also seen this in the Rohini Region and you know we also observed that we do not see much of an increase or let us say not that much increase in the south east Delhi region of you know southeast Delhi region.

(Refer Slide Time: 24:43)



Air pollution monitoring in Delhi – An environmental justice issue – A spatial prediction problem

Spatial Prediction

Source:
Delhi Pollution Control Committee
Government of NCT of Delhi

The last sort of interesting example I want to talk about is the air pollution monitoring problem of Delhi. Of course, you know Delhi suffers from an air pollution crisis really and you know different agencies, different administrative agencies have been trying to tackle this problem. How do they tackle this problem?

Well you know the first step is monitoring air pollution; you know till we know what kind how much problem till we can measure the problem that we are facing how are we supposed to you know address it? So, we have these air pollution monitoring stations which are these green dots on the table on your screen which are spread across Delhi.

Now, of course, you know these provide you the air pollution levels in Delhi at every, you know, at very high temporal frequency in the sense that you can go and measure air pollution levels every 30 minutes or even or even with lower frequency, you know, in these regions.

So, you may see that some regions are experiencing higher pollution levels than others for example, you know east Delhi tends to have higher pollution levels than let us say you know some other regions that is around let us say the airport region the south east south Delhi region and so on, but there is an environmental justice there is a very interesting environmental justice problem arises here is that you know having these monitoring stations is costly. So, we cannot have them everywhere.

But if the regions that we do not have these monitoring stations at all these gaps we do not really know what is the exact level of air pollution. Now, the question arises do the residents of New Delhi where there is where the density of air pollution monitoring stations is low to they not deserve to know the air quality that they are breathing, right.

Now this is this sort of you know sets off a problem that you know you cannot have stations everywhere, but if you do not have them everywhere you are actually sort of leaving the residents you know outside the opportunity of knowing the kind of air quality that they are breathing, right.

To solve this problem you know spatial statistics provides this tool called spatial prediction. In the sense that you know you can predict the level of air pollution at locations where you may not be directly observing air pollution you know levels during any given day or any given time right. So, it is a very interesting sort of its it opens up a very interesting you know dimension of spatial statistics or something that it brings to table which can resolve issues that are very critical like environmental justice.

Thank you.