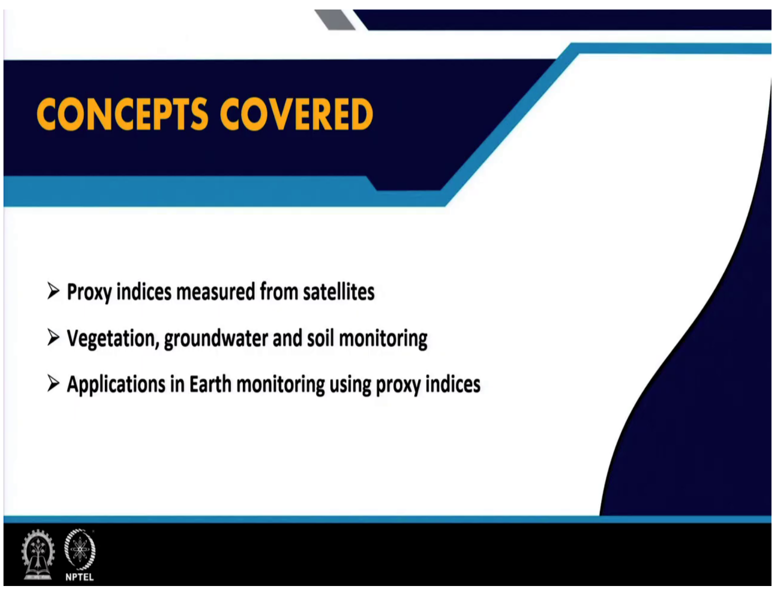
Machine Learning for Earth System Sciences
Prof. Adway Mitra
Department of Computer Science and Engineering
Centre of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

Module - 04
Machine Learning for Earth Observation Systems
Lecture - 31
Satellite Imagery as a Proxy for Geophysical Measurements

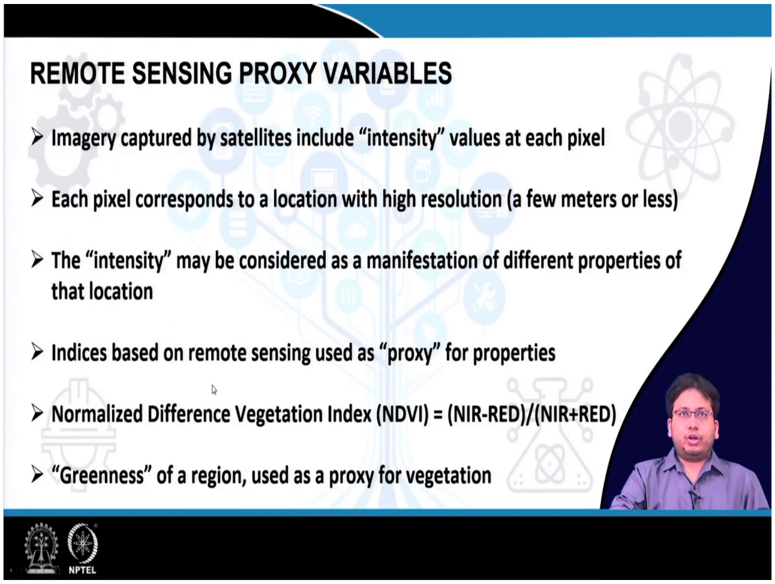
Hello everyone, welcome to Lecture 31 of this course on Machine Learning for Earth System Science. We are in Module 4 where we are looking at how machine learning can be used for various earth observation system and their applications. The topic of this lecture is Satellite Imagery as a Proxy for Geophysical Measurements.

(Refer Slide Time: 00:47)



In this lecture we are going to discuss what are these proxy indices, which can be measured from satellites, then how vegetation, groundwater, and soil monitoring can be done using like these proxy variables and machine learning measurements and like and various applications of earth monitoring using these.

(Refer Slide Time: 01:11)



So, first of all, what are these proxy variables? So, we know that like images are captured by satellites or other earth observation systems. Like basically they are images where at every pixel they contain some amount of "intensity" and these pixels as we have discussed earlier also, each pixel corresponds to a special region on the surface of the earth.

Now, like the in modern earth observation systems these images are of very high resolution and these pixels like stand for some very small region on the surface of the earth may be a few meters or even centimeters. So, it is so, basically the satellites are able to capture centimeter level or meter level information on the earth surface.

Now, the "intensity" of the pixel this can be considered as a manifestation of different properties which are present in that location, that is like the image is captured by as a result of some kind of reflectivity or something like that, but how much light or how much energy will be reflected by the earth surface at that particular region depends on what is present or what is the nature of the surface there.

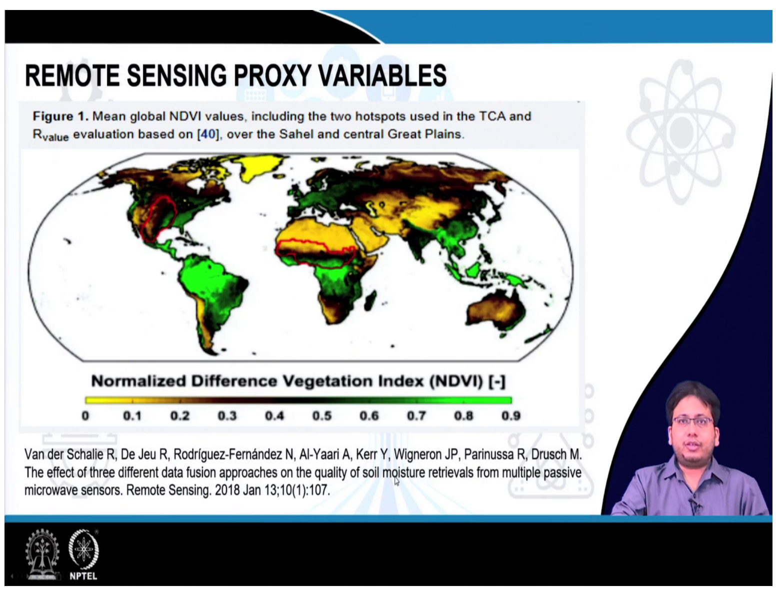
So, based on that like based on the amount of reflectivity which is or based on the amount of intensity which are obtained from a particular region in images at captured at different wavelengths it is actually possible to calculate certain indices and these indices can act as “proxy” variable. For example, a well-known index is called the NDVI or Normalized Difference Vegetation Index.

So, it is like saying like suppose we capture two images; one in NIR near infrared and the other in the RED channel of the visible range. So, at every pixel we will get in the. So, that is for the same region we will get a NIR image and we will also get a RED image.

Now, in the NIR so, at every pixel I mean at every location if we carry out this operation that is the difference between the NIR and the RED and the RED channel images divided by the their sum. So, that is known as the NDVI at that particular location. In a sense this what index captures the amount of “greenness” of that region.

$$NDVI = (NIR - RED)/(NIR + RED)$$

(Refer Slide Time: 04:00)



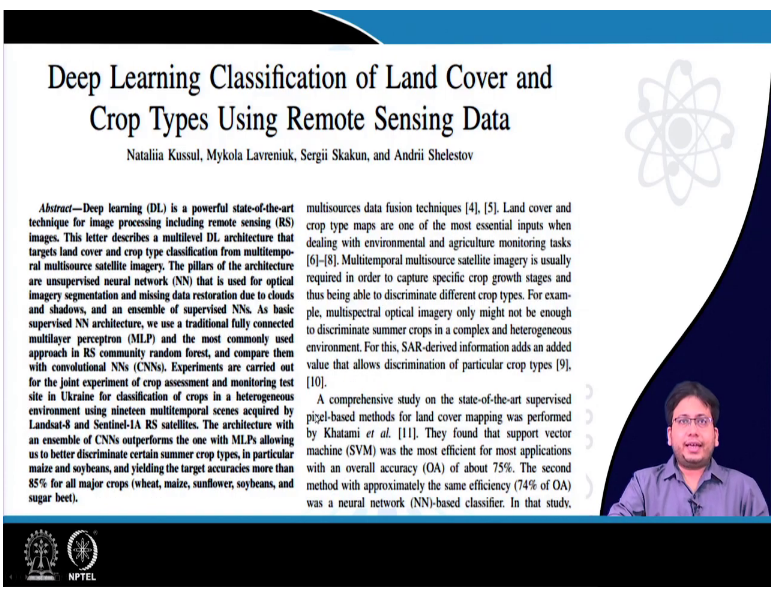
And greenness can be used as a proxy of for vegetation. So, like if we like so, if we have the observations from all over the world from which we calculate this NDVI index in the according to the formula I just mentioned, so, this is what the NDVI index looks like. So, as you can see that there are some regions which are like this kind of light green for which the NDVI value is very high, having the NDVI is because it is normalized it must lie always between 0 and 1.

And now it can happen that in some locations it is closer to 1, in some locations it is closer to 0 and in some other locations it is like in between let us say around 0.5 or something like that and it turns out that in some of the most forested regions of the world, so, let us say in the Brazil which has its huge Amazon forest it is mostly green.

So, similarly in much of Sub Saharan Africa which is also densely forested you see this strong green color. Similarly in much of these densely forested regions of South-East Asia also we see green. On the other hand in the let us say in the Sahara desert, in the deserts of Arabia and also these places of China the arid regions of China as well as even in these places like Greenland where there is no greenery like there also we see low values of NDVI.

While in other moderate like areas which are like densely populated or like built up, but have there is some amount of trees are present, but not like dense forest. So, like for example, in much of India, in various parts of Europe and Australia and so on also in the United States and Canada we see like around 0.5 value of this NDVI.

(Refer Slide Time: 06:00)



So, basically this NDVI this is the, and this index is can be used as a proxy for vegetation. There are also various other indices which can be defined in this way. So, there is something known as Leaf Area Index; LAI and various other things. So, like I leave it as a homework for you to read various or read about these various indices which might be obtained from the satellites and each of them have like can be interpreted in some particular ways.

Now, these interpretation will always have some amount of subjectivity. We cannot say like with certainty that if the like if the NDVI of a particular region is close to 1, we cannot say for with certainty that there is vegetation. It could somehow be something else also which is for which NDVI is high. Similarly, just because NDVI is say 0.4 or 0.5 like we need not be like we need not interpret it in some way. Like different values of NDVI can be attained in different regions for different for different reasons.

However, like as a like as a approximate sentence we can say that high value of NDVI generally suggests a high level of greenery. So, now, like this property is utilized like as inputs to various machine learning and deep learning algorithms nowadays and earlier like other simple threshold based mining data mining techniques to like build the maps of some geophysical variables which might be a function of the amount of vegetation in the region.

So, like one application of this approach is in this paper which deals with Classification of Land Cover and Crop Types Using Remote Sensing Data. So, a deep learning is a powerful state-of-the-art technique for image processing including remote sensing images. This letter describes a multi-level DL architecture that targets land cover and crop type classification from multitemporal multisource satellite imagery.

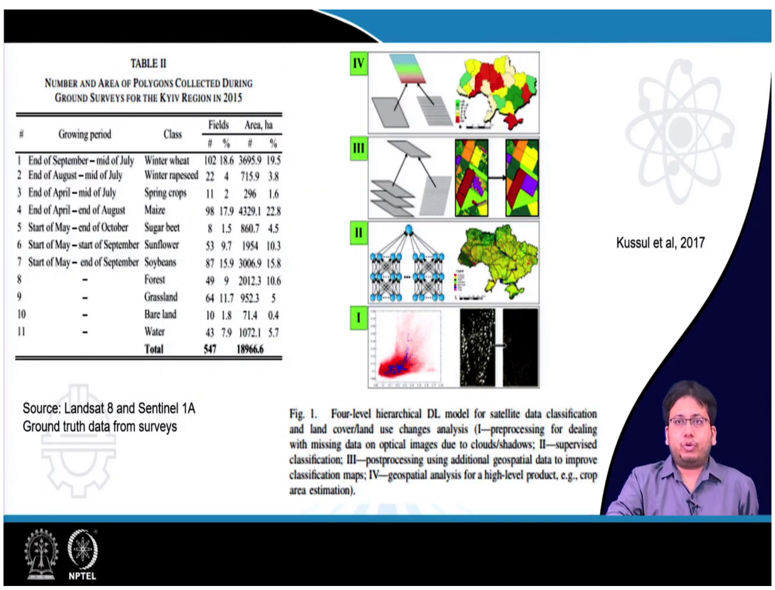
Multi-temporal means taken at different points of times and multisource means they are like taken from different satellites or from the same satellite at different like wavelengths. The pillars of the architecture are unsupervised neural network that is used for optical image segmentation and missing data restoration due to the clouds and shadows and an ensemble of supervised neural networks.

As a basic supervised neural network architecture, we use a traditional fully connected multilayer perceptron MLP and the most commonly used approach in remote sensing community random forest and compare them with convolutional neural networks. Experiments are carried out for the joint experiment of crop assessment and monitoring test site in Ukraine for classification of crops in a heterogeneous environment using nineteen multitemporal scenes acquired by Landsat- 8 and Sentinel-1A remote sensing satellites.

The architecture with an ensemble of CNNs outperforms the one with MLPs allowing us to better discriminate certain summer crop types in particular maize and soya beans and yield yielding the target accuracies more than 85 percent for all major crops. So, like here they have compared convolutional neural network with multilayer perceptrons.

So, if you remember in the previous lecture when we are talking about image fusion from multiple sources, we mentioned this multilayer perceptron being used for calibrating the satellite and sorry for calibrating ground based radar and spaceborne radar imagery and so on. So, this result suggests that it is possible that convolutional neural networks may outperform multilayer perceptrons if they are used.

(Refer Slide Time: 10:28)



So, like some of you may be interested to replicate that paper replacing the MLP with the CNN and see whether you get some better result or not. So, coming back to this paper, so, this is what they actually do. So, this is the map of Ukraine which I guess some of you may have become familiar with due to the ongoing war, but this work was of course, done in the pre like several years before.

So, now, what they have done is in this map of Ukraine they are trying to like segment it based on crop types that is such as like these colors indicate which crop is being grown in which region. So, first of all they obtain these kinds of, so these are the satellite imagery that have been captured. They have some like missing data also in the optical images due to clouds and shadows.

So, they use some kind of pre-processing techniques to cover them, then secondly, there is the classification is carried out and finally, some additional post processing data is used to improve this classification maps. So, like when we use the like image segmentation so, again.

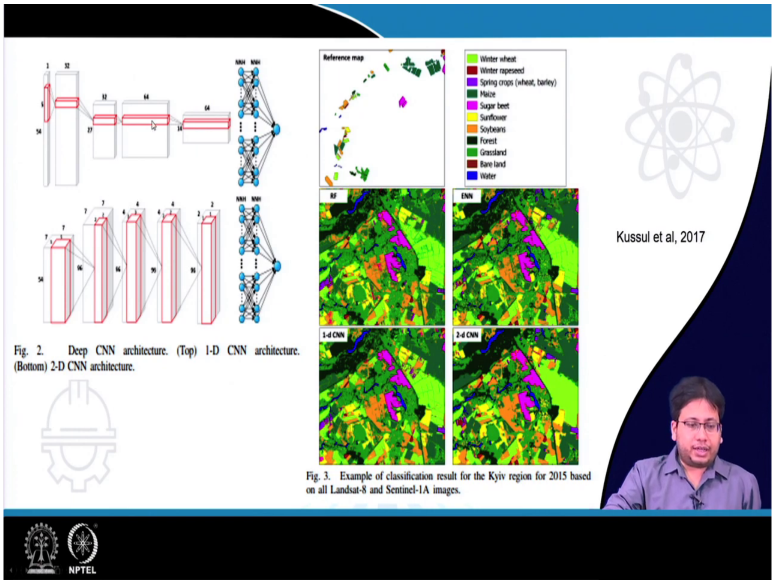
So, remember that they have not done image segmentation like in the way we discussed a couple of lecture ago by U-Net and SegNets etcetera, but rather they have like because their aim is not to like separate or segment the region the entire map according to say like the say this region is has water, this region has built-up area and so on. The instead what they have done is they are trying to do the segmentation according to the type nature of crops grown in different places.

Now, like even in the particular region it is not that throughout the year the same crop will be will be grown or any crop will be grow grown, there might be seasons. So, that is why it is in this case they need to consider a multi temporal images as they have mentioned here like they need a sequence of images on which they can carry out this kind of segmentation.

Because like if you focus on the image taken at any particular point of time some like the only a certain crops will be present. The other areas even if they are used for developing some crops in, but at that point of time that may not be the cultivation season. So, they; so, such crops may not be feasible, so that is why it is necessary to consider images from wide range of time.

So, like these are the so, the that is what they have indicated here like for which crop they need images from which period of time and accordingly they have developed the CNN architecture.

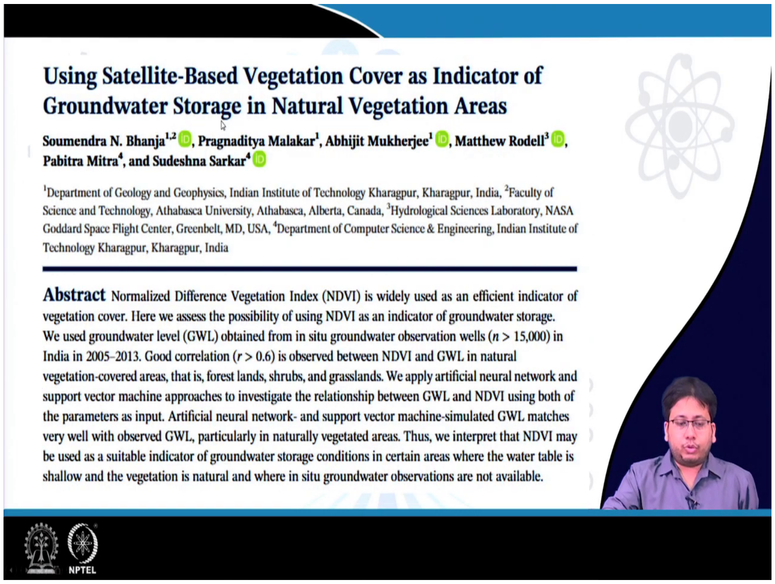
(Refer Slide Time: 13:20)



So, this is what their architecture looks like. They have actually compared two approaches. One is the 1-D CNN and the other is the 2-D CNN, but like the basic architecture wise if there is no great innovation here. Like this is pretty much the similar CNN architecture and so, but this is the classification results which they have obtained by over the multitemporal images.

Basically, the 2-D that is they do need to do this 2-D CNN because of the presence of multitemporal images. Because for every because they are dealing with not a single image, but a sequence of images that is why they need an extra layer there in this case. So, this is what the segmentation map look likes and so, these like as you can see after the segmentation these different colors stand for different crops that are grown.

(Refer Slide Time: 14:16)



And now let us come to another application. So, here the task is like they have the cell the remote sensing satellite measurements of NDVI which we just which we mentioned earlier also. Now, the aim of this work is to measure ground water level in different parts of India using this like this NDVI. Now, this might seem a little bit counter intuitive that I mean ground water is something that is underground while the satellite is several kilometers in the sky.

So, how is it possible for the satellites to measure ground water? The answer is like the ground water influences vegetation on the earth surface. The vegetation in turn is linked with NDVI that is we have already discussed how NDVI may be used as a proxy for vegetation and then the satellite can measure NDVI.So, like in this two step or three step way like we may say that satellite can be like from the satellites it is possible to get an estimate of the ground water.

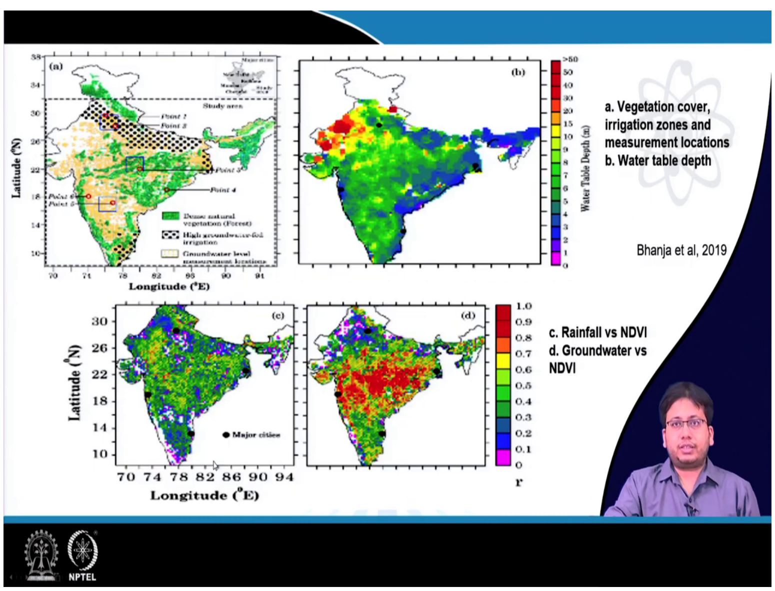
Of course, this is not uniformly like this will not always work very well as this study itself has shown because the relation between ground water and vegetation, this is not a common like it is like this is not constant. It varies from one region to another and similarly and the relation between vegetation and NDVI that is also like a somewhat dicey issue.

So, like normalized difference vegetation index or NDVI is widely used as an efficient indication for vegetation cover. Here we assess the possibility of using NDVI as an indicator for groundwater storage. We use ground water level or GWL obtained from in situ ground water observation wells like which are more than 15000 in number in India during the period 2005 to 2013.

Good correlation of more than 0.6 is observed between NDVI and groundwater level in natural vegetation or vegetation covered areas, that is forestlands, shrubs and grasslands, but not in other places like which we will discuss in a bit more detail. We apply artificial neural network and support vector machine approaches to investigate the relationship between GWL and NDVI using both the parameters as input.

Artificial intelligence, artificial neural network and support vector machine simulated GWL matches very well with the observed GWL particularly in naturally vegetated areas. Thus, we interpret that NDVI may be used as a suitable indicator of ground water storage conditions in certain areas where the water table is shallow and the vegetation is natural and where in situ observations, ground observations are not available.

(Refer Slide Time: 17:16)



So, this is this roughly belongs to that category of work where we have in situ measurements of like a certain variable and we are willing we are we wish to build like a an a continuous map of that particular variable so that I can get the map, I mean the get an estimate of the variable at any point I deserve.

But the only thing is that instead of going for something like say Gaussian process regression or that kind of things which we had discussed in some of our earlier papers like here they are trying to do it on the basis of remote sensing and satellite imagery. So, this is like on the map of India these are the like this is basically showing first of all the vegetation cover, the irrigation zones and the measurement locations.

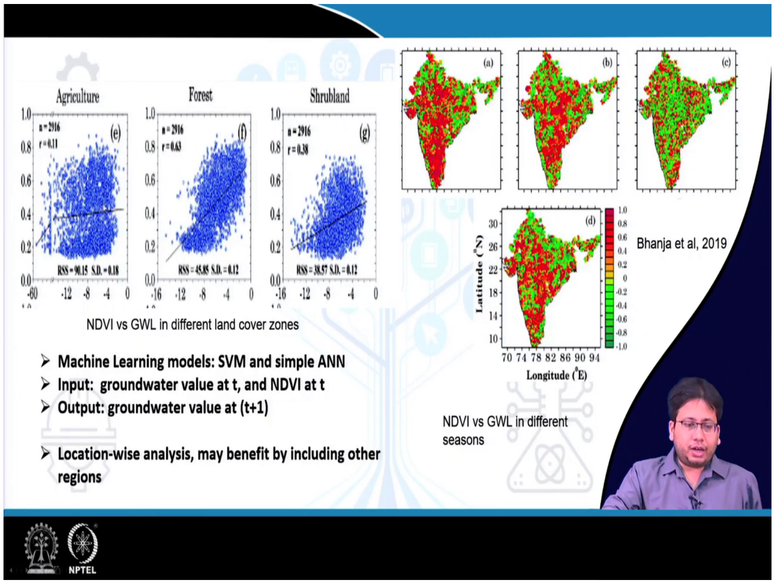
So, these red dots as you can see, these are some of the locations where in situ measurements are available. In fact, these are only some of them. There are actually as they have indicated here there are some more than 15000 places mostly wells where the ground water can be measured.

Now, these black dots what they indicate these are the places where like the irrigation for farming is done mostly by drawing water from the ground by say something like tube wells or something like that. In other regions of India when like irrigation is done for agriculture it is either rain fed or maybe from other sources like say rivers and so on.

But in these places like there are there is a system of drawing water from the ground and then like pouring that water over the crops and so on and now we will see how that becomes significant in this study. So, this and this is like a measure of the water table in different parts of the country. This is obtained by like based on a few in situ measurements and followed by some kind of spatial interpolation, this is this approximate map is obtained.

Here we see the correlation coefficient between rainfall and NDVI. And as you can see that in many places there is only weak correlation. It is either like quite low that is say 0.1, 0.2 something like that or moderate like in the range of 0.5, but never above that. On the other hand here like this map shows the correlation coefficient between ground water and NDVI and here we see like in at least in some places the correlation is actually very strong. So, that is we see quite like this red covers.

(Refer Slide Time: 20:03)



And now here like there are some further studies which this paper has done like in different seasons as well as in different types of land cover zones they have carried out the correlation or they have computed the correlation between the ground water level along the x axis and the value of the NDVI. So, they have done it in agricultural regions, that is where there is a lot of cropping, then in forest regions as well as shrublands.

So, as the correlation coefficient suggests in forest regions like which are present in much of Central India like we see like a near like a quite good coefficient of 0.63 which in the earth sciences can be considered as reasonably high like there is. So, there is almost; there is reasonably high correlation between groundwater level and NDVI which means the greener or more the ground water greener that area is.

This is true for forested regions. If you consider shrubland the relation is a bit weaker, but nevertheless some relation is there, that is more the or higher the water level more green will the shrubland look, while in case of agricultural region this is not at all the case. There is a very poor correlation between the two. That is because like as I mentioned earlier also, the agriculture much of agriculture depends on this irrigation.

So, irrigation can be either rain fed, ground fed or river fed. So, in those places like especially in large parts of North India there like where irrigation is water irrigation water is drawn from the ground. The NDVI value might be quite high because it is an irrigated area.

So, like from the top like it might appear green, but due to like so much usage of groundwater over these years, the groundwater level has significantly depleted in many of these places. Now, this is something which is missed by the NDVI analysis. So, like in the these agricultural regions particularly in the in much of North India the like NDVI can be deceptive as a proxy if we use it as a proxy for ground water that is what this paper finds out.

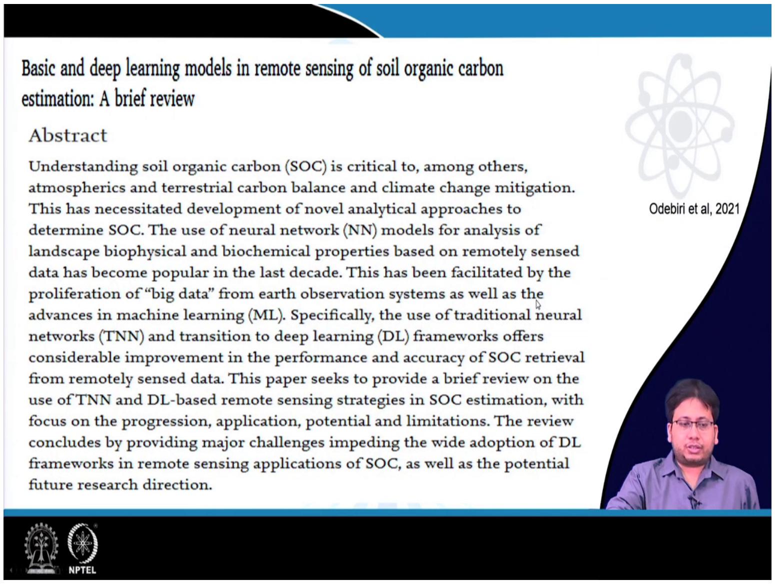
So, what they try to do? So, on the basis of this correlation study what they try to do is to build some kind of a machine learning model because we see our reasonably linear relation. So, instead of going for any complex or non-linear model like say CNNs and so on like here they use a simple support vector machine and simple artificial neural network, where the input is like the ground water value at that location at time t as well as it is NDVI at time t.

Now using that, they are trying to predict the groundwater value at time t plus 1. So, like the ground water as we know like it has a certain amount of seasonality, it have also has a certain amount of this. Like we can say it has a certain amount of memory if you want to call it that way that is the ground water is not going to change in a matter of 1 or 2 days.

Like, if it is already high it will remain high, if it is already low it will remain low, it might increase or decrease a little bit, but there will not be any significant change. So, that is why it is important to consider the ground water level at the previous time point also to make a prediction about the future, but the NDVI acts as like you can say acts as a dynamic component for this prediction task.

So, in this work they have done location based analysis that is it is like for every location they have like used this model separately, but like instead of that if they try to use the like neighbouring locations also in making this kind of a prediction like it may increase or this performance may improve due to the considering the like neighbourhood effects.

(Refer Slide Time: 24:18)



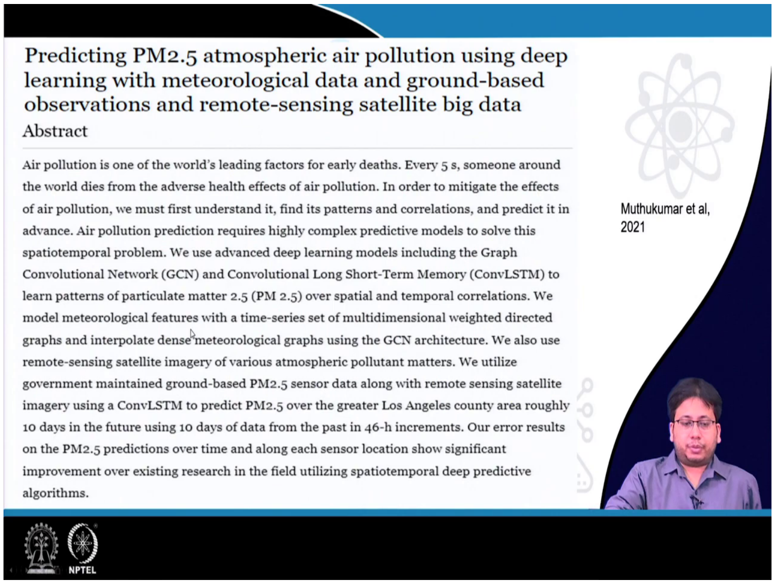
So, now here we move on to another application of these satellite based indices. So, here the task is remote sensing for soil organic carbon estimation. Understanding soil organic carbon or SOC is critical to, among other things, atmospherics and terrestrial carbon balance and climate change mitigation.

This has necessitated the development of novel analytical approaches to determine the soil organic content. The use of neural networks for analysis of landscape biophysical and biochemical properties based on remotely sensed data has become popular in the last decade.

This has been facilitated by the proliferation of "big data" from earth observation systems as well as advances in machine learning. Specifically, the use of traditional neural networks and transition to deep learning frameworks offers considerable improvement in the performance and accuracy of SOC retrieval from remotely sensed data.

This paper seeks to provide a brief review on the use of TNN by TNN we mean traditional neural networks and deep learning based remote sensing strategies in SOC estimation, with focus on progression, application, potential and limitations and so on. So, like we will not go into the details of this paper.

(Refer Slide Time: 25:46)



So, but basically, this is like about measurement of soil organic carbon and we skip to one more paper. Here the aim is the prediction of $PM_{2.5}$. So, that we that we know is like in the measure of the air quality or like it is a air pollutant atmospheric. So, predicting $PM_{2.5}$ atmospheric air pollution using deep learning with meteorological data and ground-based observations and remote-sensing satellite big data.

Air pollution is one of the world's leading factors for early deaths. Every 5 seconds someone around the world dies from the adverse health effects of air pollution. In order to mitigate the effects of air pollution, we must first understand it find its patterns and correlations and predict it in advance. Air pollution prediction requires highly complex predictive models to solve this spatiotemporal problem.

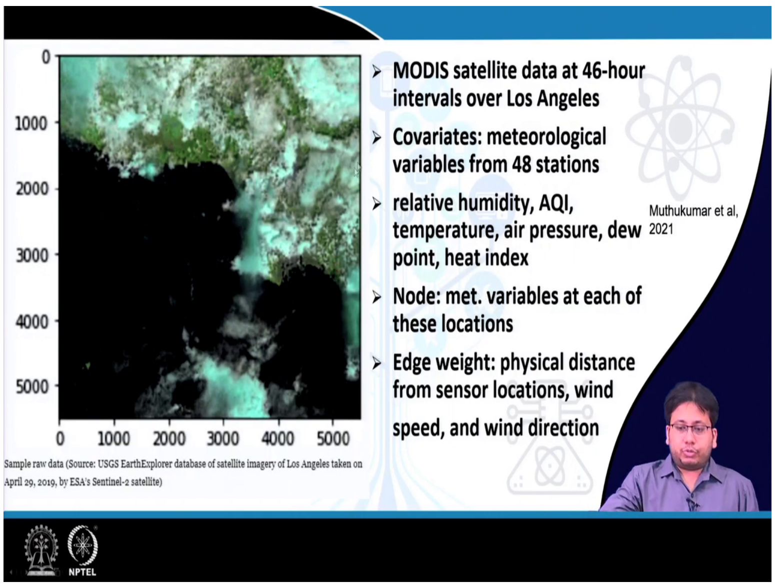
We use advanced deep learning models using graph convolutional network or which is GCN and convolutional long short-term memory ConvLSTM to learn patterns of particulate matter or $PM_{2.5}$ over spatial and temporal correlations. We model meteorological features with a time-series of dimensional of multi-dimensional weighted directed graphs and interpolate dense meteorological graphs using the GCN architecture.

So, this is like somewhat new approach which this paper has adopted in the sense that like the time series of these different variables which they are studying they have represented it as a graph. And we will I will show you in the following slides how exactly that representation has happened, but that and their aim is at in like. So, they have the graphs at different observation station.

Now their aim is to is like estimate that graph at other locations also for which they have used this graph convolutional network, which is a like you can say it is a CNN only, but it like in the in this case the different the input is like instead of a vector or a matrix it is something like a graph and like. So, like we can say the nodes of the of that graph as well as their edge structure this is somehow expressed as some kind of an embedding in the neural network.

We also use remote sensing satellite imagery of various atmospheric pollutant matters. We utilize government maintained ground-based $PM_{2.5}$ sensor data along with remote sensing satellite imagery using a ConvLSTM to predict $PM_{2.5}$ over the greater Los Angeles county area roughly 10 days in the future using 10 days of data from the past in 46 hour increments.

(Refer Slide Time: 28:52)



So, basically the source of their data comes from the MODIS satellite which is at 40 like which like we can say it is available for a particular region over the Los Angeles at 46 hours intervals.

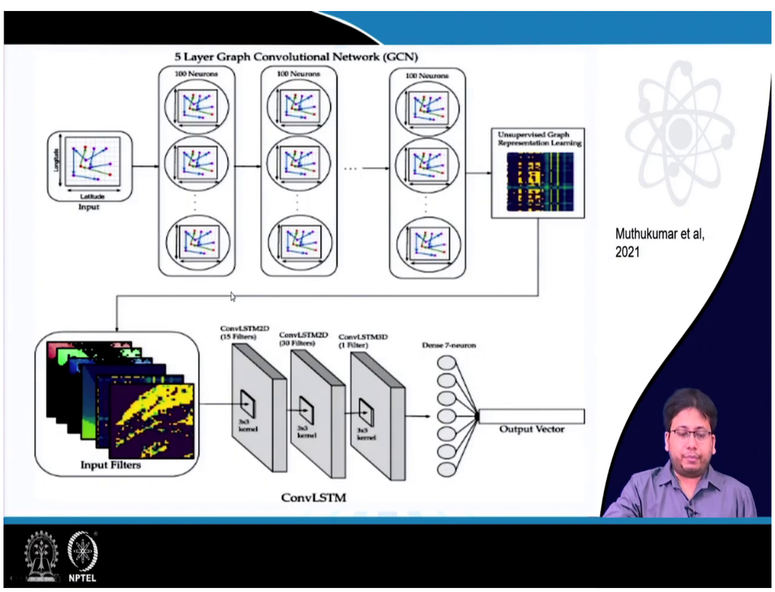
So, like they also apart from this satellite imagery like an example of which you are seeing here it is also like they also can have various covariates which are meteorological variables measured at 48 stations.

So, you can say that these are the in situ measurements. So, these covariates are the relative humidity, the AQI air quality index, temperature, air pressure, dew point, heat index, etcetera. Now, from all these things they plan as they have mentioned they plan to build the map of PM2.5 in the like in for future the for the next 10 days in future.

So, now what they do as indicated earlier, what they want to do is, they want to build a graph. Each node in the graph represents the; these met variables at each of these locations and the a the. So, whenever there are two nodes in a graph they are all they have to be connected by an edge.

And the weights of these edges like depend on the physical distance from their locations on the wind speed, wind direction and various other things which they have mentioned.

(Refer Slide Time: 30:17)



So, like as I said now at a different locations the input is this kind of a graph. And so, this is the graph convolutional neural network the GCNN. So, like all the like as this figure suggests, so, like every hidden node in this CNN like is actually some kind of a representation or like an or as they say an embedding of the graph on which the different operations are carried out.

So, that creates some kind of an like an unsupervised graph representation that is these graphs that have been provided as input like it is somehow like we get some kind of a like a representation. And now when the input comes in like from the remote sensings, imagery and so on, I mean primarily the remote sensing imagery, so, this graph is obtained on the basis of these meteorological variables which already mentioned.

Now, these the this graph is somehow like is merged with the remote sensing imagery as they come in at 46 hours interval and this merging is achieved with the help of a convolutional LSTM. So, what how exactly a convolutional LSTM works? We have discussed a bit in one of our earlier lectures and we will again discuss it in more detail in the next lecture.

But finally, we get the output which is the like at any particular location which has been specified here as the input like it will it is going to be; the output is going to be the $PM_{2.5}$ measurements for the next 10 days. So, like note that here they are using convLSTM because it is something like a LSTM because it is a sequence of maps for the last few days or so and then like which is obtained by merging the graphs and also by from the remote sensing imagery.

So, we like, we get a sequence of this so that is why LSTM and also Conv because these are all like spatial maps. It is not a particular location, but it is a map from over the entire region.

(Refer Slide Time: 32:42)

So, that brings us to the end of this lecture. So, these are the references of the papers which we discussed in this lecture. So, basically the point to be taken away is that is the different variables such as air quality, soil organic carbon component, groundwater etcetera.

The while these cannot be measured directly from satellites, but there are different proxies; proxy variables like in some cases the indices like NDVI, in some cases maybe something else or other indices like they like which have may have some loose relation between with the quantity of our interest.

So, that kind of relation like it is possible, like it may be possible under certain circumstances to parameterize using some kind of a machine learning model and hence use the remote sensing imagery to obtain estimates of that variable. So, that brings us to the end of this lecture.

Thank you and in the next lecture we will see one more application of this remote sensing imagery using machine learning. So, till then bye.