Machine Learning for Soil and Crop Management
Professor Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology Kharagpur
Lecture 48
Digital Soil Mapping – General Overview (Contd.)

Welcome friends to this 3rd lecture of week 10 of NPTEL online certification course of Machine Learning for Soil and Crop Management. In this week, we are talking about general overview of digital soil mapping. In our first two lectures, we have discussed the basic concept of digital soil mapping, why it is important, what are its uses, and also we have seen the different types of applications.

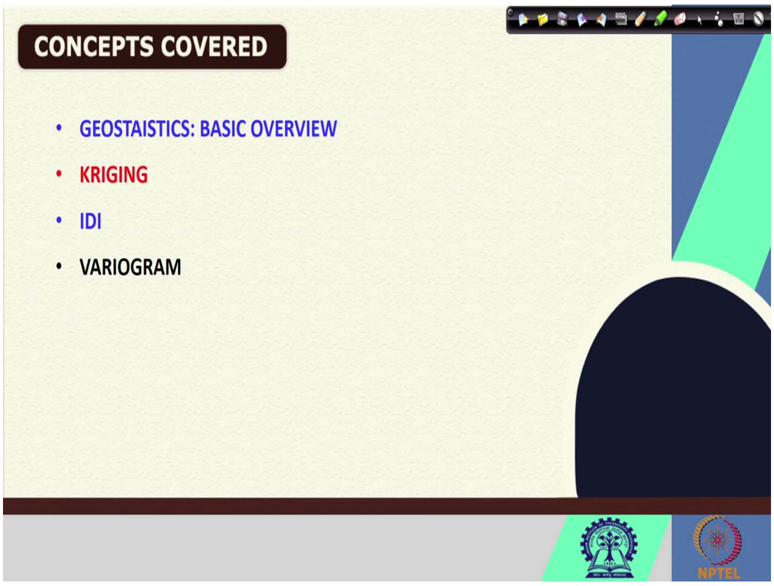
And also different models for soil formation starting with the Jenny's model of soil formation. We have also discussed the SCORPAN plus e model given by McBratney et all in 2003, which provides the fundamental backbone of digital soil mapping and why it is called digital soil mapping we have also discussed; different types of digital sources of data and how we can handle those data we have discussed.

We have seen different covariance data DEM and different types of terrain parameters, climatic variables, which you can use for producing the digital soil map which we have also discussed. We have discussed the workflow in DSM, we have also discussed the principle of DSM and also we have discussed how we can take management and policy decision using digital soil mapping.

Apart from that we have also discussed the basics of GIS. Why GIS is important nowadays and also what are the different types of data structure in GIS, we have seen. We have also discuss the different types of coordinate reference systems starting with geographic coordinate system, also we have discussed UTM or Universal Transverse Mercator. And we have also discussed what is datum.

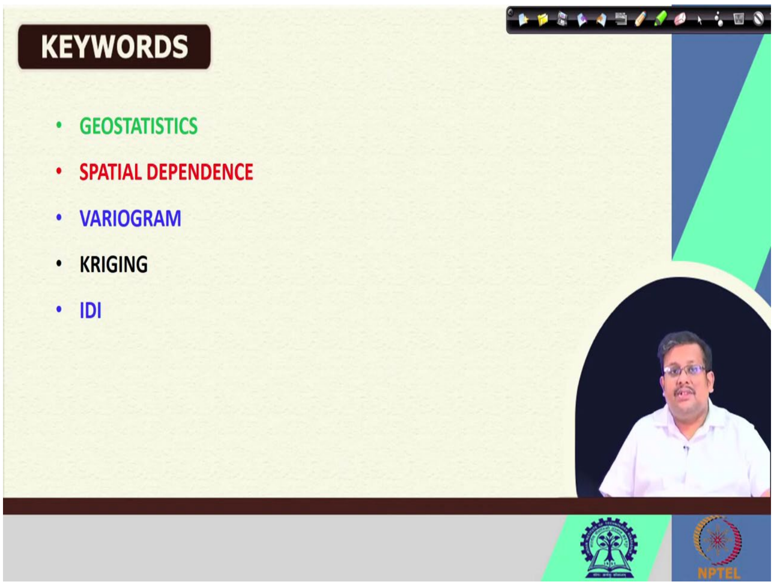
So, today in this lecture, we are going to discuss a very important topic which is extensively used in digital soil mapping or rather I would say that digital soil mapping is based on this particular discipline, sub-discipline of statistics, this is called geo statistics.

(Refer Slide Time: 2:52)



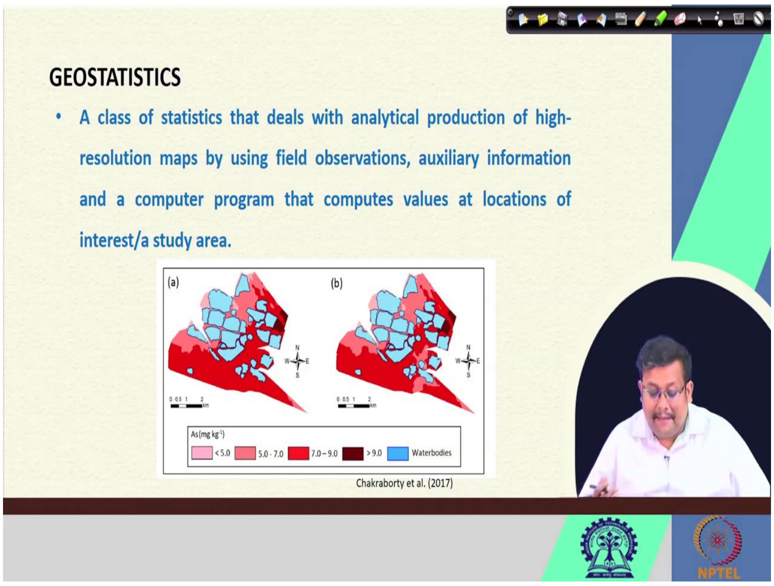
So, in this lecture we are going to have the overview of this concepts, we are going to have basic overview of geo statistics. Of course, the time is limited so I will try to cover the basic aspects of geo statistics. In this lecture, we will be learning about Kriging interpolation, then inverse distance interpolation and also we are going to talk about the Variogram and why it is important.

(Refer Slide Time: 3:19)



These are the keywords of this lecture, geo statistics, spatial dependence, Variogram, Kriging and IDI.

(Refer Slide Time: 3:27)



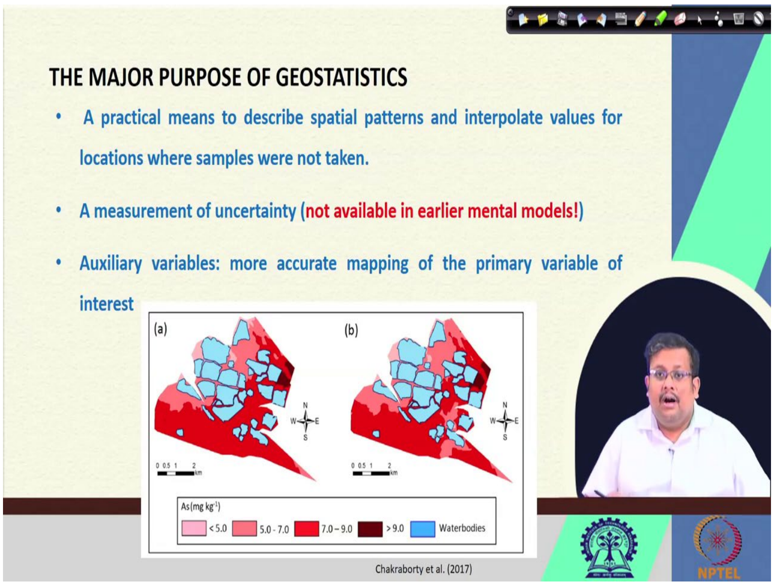
So, let us define what is geo statistics. So, it is a class of statistics that deals with analytical production of high resolution maps by using different field observations, auxiliary information and a computer program that computes values at location of interest or a study area.

So, whenever our resources are limited, and there is not possible to collect enough number of samples from the whole study area or the location of interest, then we take the help of geo statistics who are doing different types of interpolation, what is interpolation is prediction and map of properties of in the context of geo statistics, the definition of interpolation is prediction of properties soil properties at locations where there was no preliminary sample.

So, in other words interpolation helps in predicting the properties at unsample location as you can see here, these are two interpolated maps of soil arsenic concentration, which we have created in one of our previous research and published in the Journal of Geo derma. Here we can see the distribution, spatial distribution of the soil properties of course, these were made by taking only hundreds of samples.

So, using those samples we have, we have produced this map. So, that shows that shows the actual application of geo statistics to predict the property at unsample locations also.

(Refer Slide Time: 5:30)



So, what are the major purposes of geo statistics? There are a couple first of all, it gives you the practical mean to describe the spatial patterns and interpolate values for location where samples were not taken of course, this is a major rational behind this interpolation spatial interpolation, apart from that, it gives you at the same time, it gives you a measurement of uncertainity.

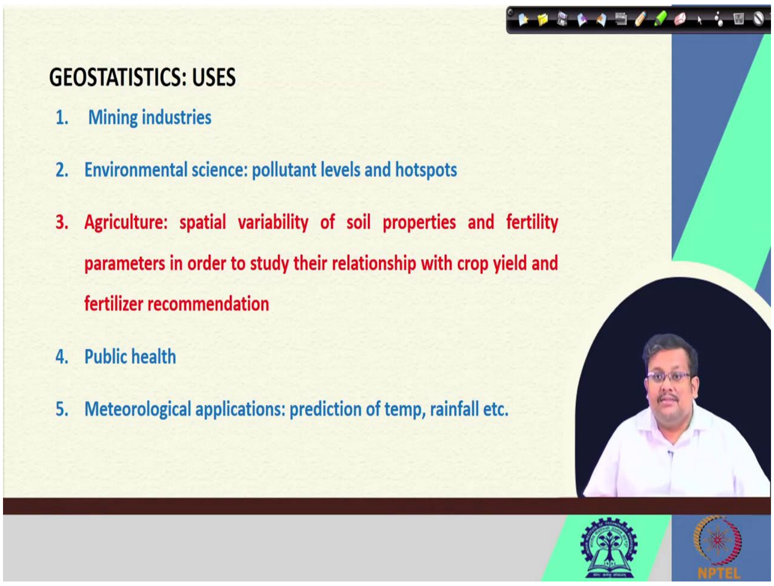
Now, generally, before the advent of geo statistics in the soil science domain, we have seen that the experts are developing their maps by their observation and by the knowledge of correlation with other auxiliary properties. So, these are called the maps made by mental models. However, these models were not able to generate a measure of uncertainty.

So, you do not have any idea about how accurate your mapping is so, but geo statistic gives you not only the map using the predicted values but side by side, it gives you the map of uncertainty it is also so, that you can have enough confidence while interpreting the map using geo statistics.

Now, also we have seen that when there are several auxiliary variables which you are using for producing the geo statistical interpolation that helps you for getting better accurate map of the primary variable of interest because of their correlation with the primary variable of interest. So, we have already seen what are those covariates we generally use in geo statistics in DSM.

We have seen the climatic variables, we have seen the terrain variables, we have seen the remote sensing variables, all these are used as auxiliary variables in the geo statistical interpolation to get more we generally use them to get higher prediction accuracy and subsequent better mapping of the primary a property of interest.

(Refer Slide Time: 7:50)



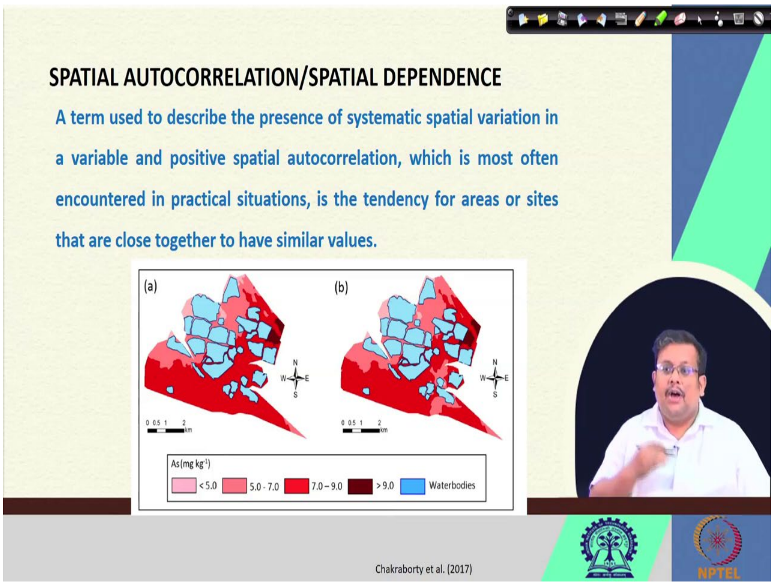
So, what are the uses? Of course the geo statistical application first was initiated in mining industries, then environmental science for pollutant level pollutant level and hotspot identification. And it began and lately in last couple of decades, people have started using this for agricultural operation.

For example, in the spatial variability of the soil properties and fertility parameters, if you are interested to see what is the spatial variability of fertility parameters in your in your field agricultural field, then you have to take the help of geo statistics. And also if you want to make a correlation with the spatial variability of soil fertility properties with your crop yield, then also and also you want to make some fertilizer recommendation based on those mapping you have to take the help of geo statistics.

Of course, apart from that public health and metrological applications like prediction of temperature, rainfall, all these are dependent on some geo statistical approaches. So, you can see that agricultural indicates one of the major area of application of geo statistics. Now a days you the advent of different types of machine learning different types of machine learning approaches, these geo statistical approach or geo statistical interpolations are getting improved each and every day.

And new methods are coming and with the help of these hyperspectral resolution hyperspectral remote sensing high resolution data, we are getting higher resolution of soil properties map. So, using in combination with different types of machine learning algorithms. So, these are some of the applications of geo statistics.

(Refer Slide Time: 9:46)



And but before we discuss the basics of geo statistics. Let us understand what is spatial autocorrelation or spatial dependence is a very important term and this is the backbone or I would say, this is the major motivation behind geo statistics areas application. Why we could do geo statistics, this is because of spatial interpolation sorry spatial autocorrelation or spatial dependence.

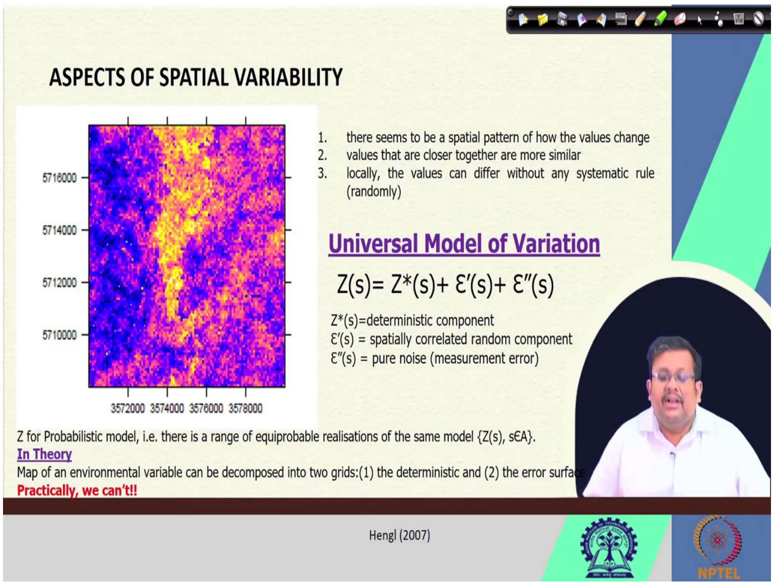
Now, this spatial autocorrelation or spatial dependence is a term that is used to describe the presence of systematic spatial variation in a variable and positive spatial autocorrelation which is most often encountered in practical situation is the tendency for areas or sides that are close together to have similar values if you can see these maps, one thing is for sure, you can understand that the locations which are close by are showing similar properties or similar levels of properties.

And those areas which are far apart are having some differences and also there is an some abrupt changes of such properties. So, you can see that this is called spatial autocorrelation that means the observations at a particular distance are not independent to each other they are closely related to each other. And this is called spatial dependence or spatial autocorrelation.

In real life situation also in the agricultural field, you will see that the soil properties which are which are measured from the nearby samples, they match with each other. However, when you take a soil sample far apart from the primary location of soil sample, you will see that these two samples may differ in their properties.

So, one thing is clear that the data is some correlation of the soil properties based on the spatial distance between the sample point. So, this is called spatial dependence or spatial autocorrelation. Now, geo statistics exploits these phenomena to map the variability spatial variability of the soil properties. So, all these maps are being developed by keeping in mind that phenomenon of spatial dependence using some defined models which we are going to discuss.

(Refer Slide Time: 12:24)



Now, let us see what are the aspects of spatial variability if we can see here in this is the spatial variability map of soil property you can see there seems to have a spatial pattern of how the value changes, values that are close together are more similar just like I have showed you in my previous slide. And locally the values can differ without any systematic rule or randomly so, we can see all these things.

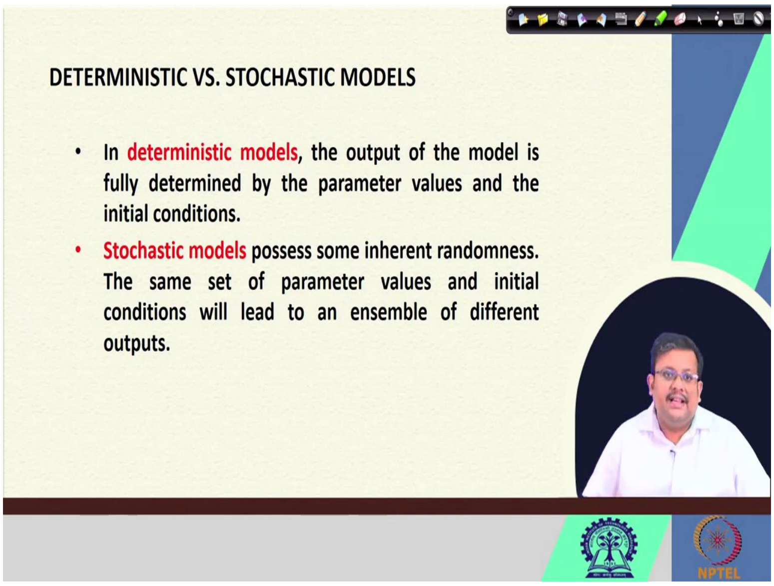
So, the variation or spatial variation of soil properties can be explained in terms of universal model of variation. Now, what is universal model of variation universal model of variation is says that the property at a given point will be combination of deterministic component which is denoted by these Z star and then especially correlated random component which is eta dash and then eta double dash is basically the pure noise or this is basically the measurement error.

So, Z is for the probabilistic model that means, there is a range of equal probable realization of the same model. So, we can see that the universal model of variation says that a property of the given place is basically a combination of the output from a deterministic model and a stochastic model and beyond noise.

Now, from the point of view of geo statistics and point of view of a mapper we are more focused towards mapping this deterministic component and specially correlated stochastic component. Because pure noise which could be due to measurement error, this is not our primary focus area because this may might be the beyond the control of the mapper.

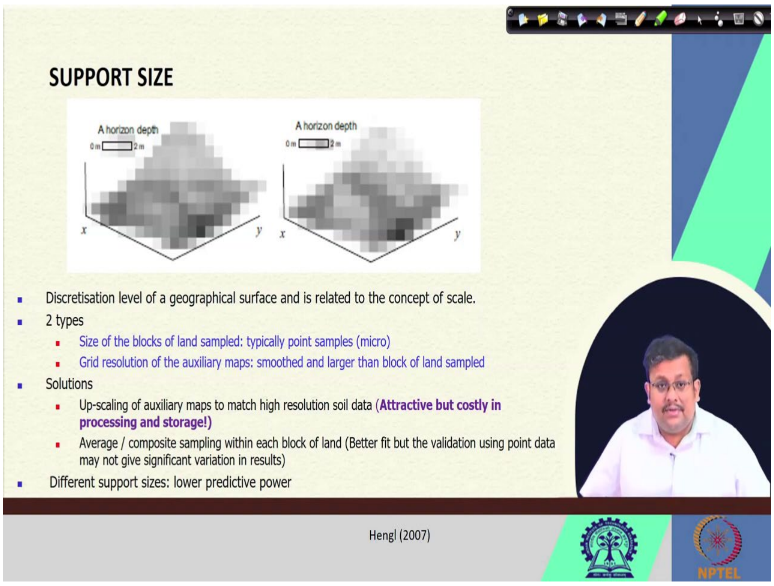
But at the same time so, the major variation comes from this deterministic component as well as the spatially correlated random component. Now, in theory map of an environmental variable can be decomposed into two grids one is deterministic and other is error surface but practically it is very difficult to decompose into those two grids.

(Refer Slide Time: 14:41)



Now, let us discuss what is deterministic model and how it differs from a stochastic model. Now, in the deterministic model the output of the model is fully determined by the parameter values and the initial condition. So, it is more or less fixed however, stochastic model is more, more random. So, stochastic models poses some inherent randomness the same set of parameter values and initial conditions will lead to an ensemble of different outputs. So, this is the major difference between deterministic model as well as the stochastic model.

(Refer Slide Time: 15:21)



Now, let us see what is support size, support size basically shows the discretization level of a geographical surface and it is a concept of scale it based on the concept of scale. Because you are developing what you are downloading the data at different scales you can download a raster file in 90 meter by 90 resolution and two meter resolution and you can download another raster file of covariates with 30 meters by 30 meter resolution.

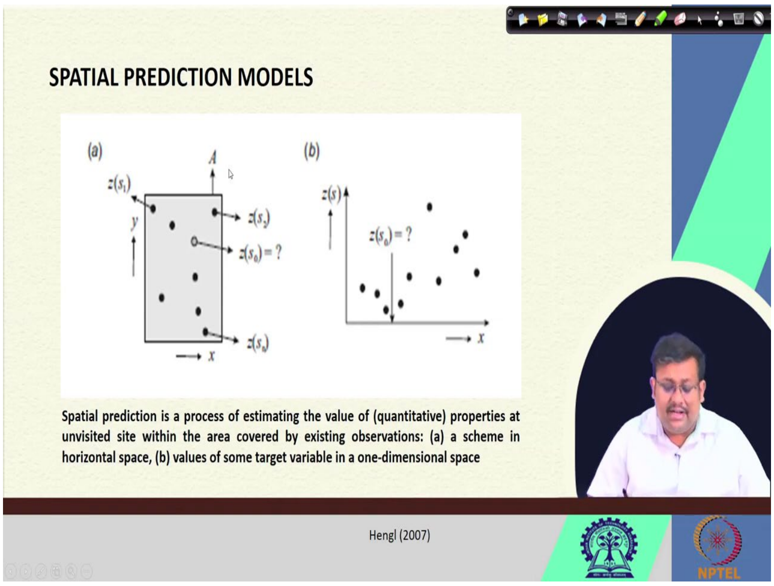
So, this 30 meter by 30 meter resolution or 90 meter by 90 meter resolution is known as the support size. So, support size varies from one layer to another layer. And it is very important that to improve the prediction accuracy, you need to homogenize all the support size by using several methods, most of the cases the samples which you take from the soil are point sample.

So, that does not matches with the support size of any other covariate which is having higher support size suppose 90 meter by 90 meter. So, in that case, we have to deal with different types of we need to do some modification for example either we can upscale the auxiliary map to match the high resolution data this could be very attractive, but it is costly in processing storage.

So, you can upscale the auxiliary map which is originally 90 meters and by 90 meter resolution to say suppose 5 meter by 5 meter resolution to match with the soil sampling point. But this would hugely enhance the cost of processing as well as the storage size or the other way of dealing with is average or make some composite sample within each block of land and so, that can give better feed.

But at the same time validation using point data may not give the significant variation in the results. So, this is why we generally suggest that you should homogenize the data for matching the support size which will help you to increase the predictive power otherwise, if there are different support sizes that will decrease the predictive power.

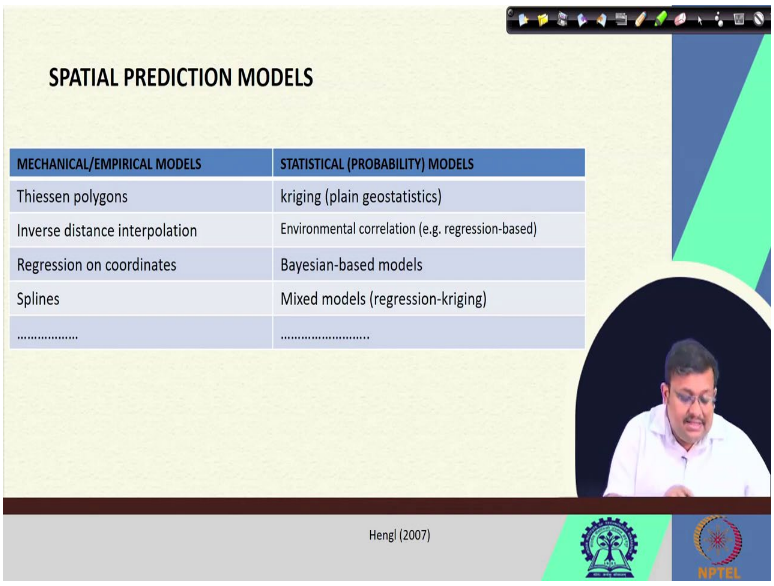
(Refer Slide Time: 17:40)



Now, if you see the spatial prediction model, this is a process of estimating the values of quantity of quantitative values of course, at unvisited side within the area covered by existing observations. So, these are our area of interest which is displayed to in two dimensional place x and y bounded by. So, here these z s1 s2 and all these observations the solid dot black dots are showing the original sampling points and suppose you want to predict the value at an uncertain location.

So, here you have to take the help of spatial interpolation. So, this is the scheme of in the horizontal space and this is the values of some target variable in a one dimensional base space. So, the and this the whole area of interest is noted by this capital A. So, we use different types of spatial prediction to predict that property of at sample is zero point. So, what are those what are those methods?

(Refer Slide Time: 18:49)

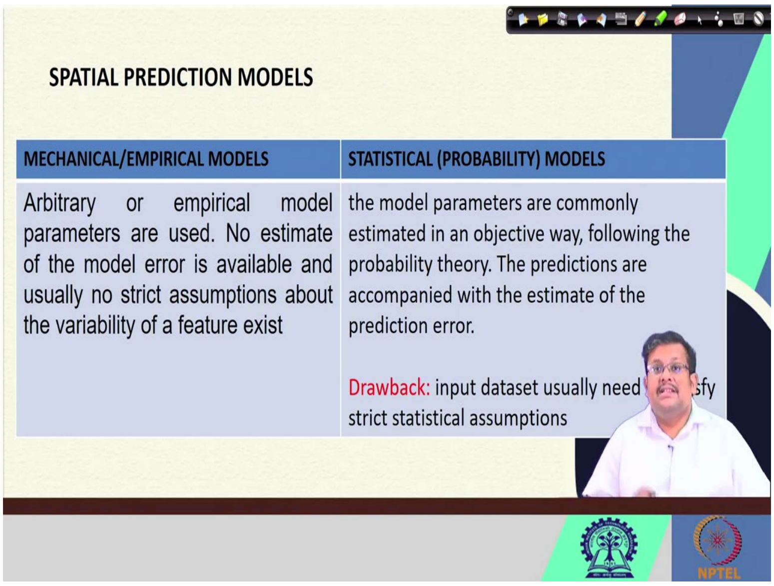


So, these methods can be called generally as a spatial prediction models. So, there are two types of spatial prediction model one is called mechanical or empirical model and second one is called statistical or probability based model. Now, some examples of mechanical or empirical models are Thiessen polygons, inverse distance interpolation, regression coordinates splines.

However, in case of probabilistic model like Kriging, which is the most widely used method and also which is also very synonymous with the geo statistics. And also you can do some environmental correlation which is will be regression based Bayesian base models and mixed models like regression Kriging.

So, we are going to learn this regression tuning in our upcoming lectures but at the same at this time, let us focus mostly on these two one is inverse distance interpolation and Kriging.

(Refer Slide Time: 19:54)

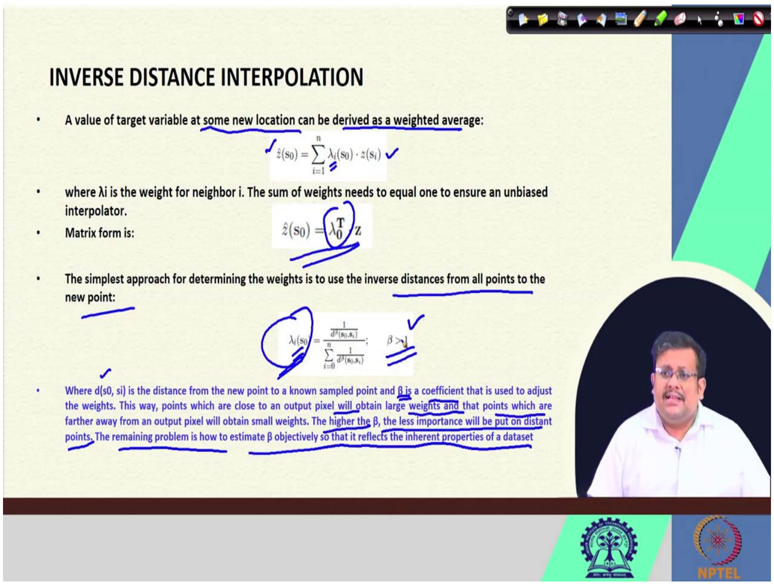


So, what is the difference fundamental difference between these mechanical model and statistics model in case of mechanical models, they are arbitrary or they are our empirical model parameters we use and no estimate of the model error is available, and usually no strict assumption about the variability of a feature exist.

So, this is called the mechanical or empirical model So, arbitrary empirical model parameters are used, when we will discuss how to create these, these models we will we will just we I will show you, but here just mentioned just understand that the spatial dependence cannot be measured objectively in this kind of mechanical or empirical model we are just assuming some arbitrary values.

And then it is a kind of a trial and error method to see which one is giving you the best interpolation. However, in case of statistical or probabilistic model, the model parameters are commonly estimated in an objective way following the probability theory and the predictors are accompanied with the estimate of the prediction error. So, input data set usually needs to satisfy but there is of course a drawback because in case of statistical probability models these input data set usually need to satisfy strict statistical assumptions.

(Refer Slide Time: 21:22)



So, let us first discuss one example each we will start with the inverse distance interpolation. Now, in case of inverse distance interpolation a value of the target variable at some new location suppose there is s0 can be derived using a weighted average with the sample with the already sampled points.

So, here we can see that this is the predicted values that s0. And this is how we are going to calculate the value at this point by multiplying this weight of the neighbors and also with the values of the neighbors. So, here these lambda i shows the weight of the neighbor i and the sum of the weights needs to be equal to 1 to ensure an unbiased interpolation.

So, this is the for general form of IDI and the matrix form you can see here this is the matrix from so, the simplest approach so, this is lambda transpose of lambda 0 multiplied by Z. So, the simplest approach for determining the weights is to use the inverse distance from all the points to the new point.

So, basically, it also in case of inverse distance interpolation, we give more weightage to the points which are close by and less weightage to the points which are far apart from the location of interest. So, here, this for executing the inverse distance interpolation, the idea is you give more weight, you assign some arbitrary values to give more weight to the nearby points and less weight to the far to the distance points.

So, this is this lambda how we calculate this lambda. So, this lambda is calculated by using this formula 1 by d beta s0 is 1 and then summation of i equal to 0 to 1 by d beta and then s0 s1 and where beta equal to greater than 1. So, here this distance s0 is i is basically is the distance from the new point to a known sample points.

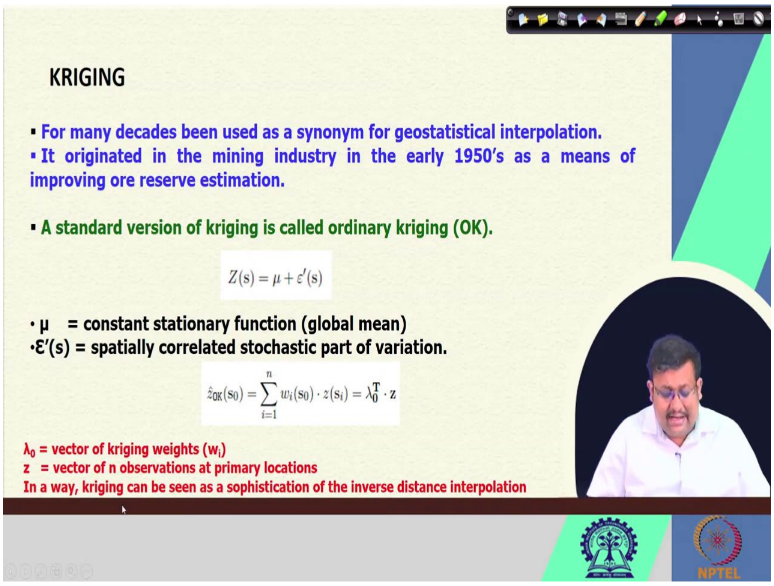
So, basically we calculate the distance from the new point to a known sample points and beta is the coefficient that is used to adjust the weights. So, beta is arbitrarily given. So, this way, So, the points which are close to an output pixel will obtain large weights and those points which are farther from the output pixel will opt in small weights just like I told you.

So, the higher the values of the beta you can see the higher the values of the beta, the less importance will be put to on distance points, and the remaining problem is how to estimate this beta objectively So, that it can reflect the inherent properties of the data set.

So, you can see here in this calculation, these weight is being calculated by using this formula, here d s0 1 is or si is the distance from that unsample location to a nearby sample location or any sample location and this beta is the weight. So, higher the value of the beta by this formula we can see that lower will be its weight. So, lower importance will be given to the distance points and higher importance will be given to the points which are close enough.

So, these beta values are greater than 1 we can try with different beta values and see which one is gives us the which reflects accurately the inherent properties of the data. So, you can see here the beta values we are assigning is not collect is not calculated in an objective way. So, this is the problem with inverse distance interpolation.

(Refer Slide Time: 25:39)



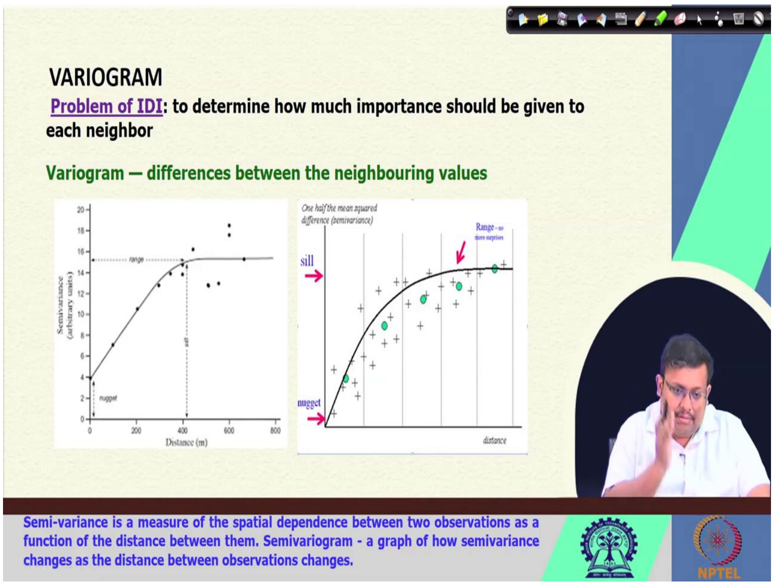
Now, the Kriging is an advancement from this inverse distance interpolation in case of Kriging for we synonymously as I have told you that we generally use this screening synonymously with geo statistical interpolation it was originated in in 1950s as a means of improving or reserve estimation in the mining industries.

So, a standard version of the Kriging is known as the ordinary Kriging the expression is this of course, you can see that, the value at any given location is basically composed of two components, one is constant stationary mean or constant stationary function or global mean and this is the spatially correlated stochastic part of variation.

So, global mean plus stochastic variation combined together to give the property at any given location. So, the formula of Kriging shows that it is same just like we have seen the in case of IDI also. So, it is a weight of any given points weight of an unsample location from a sample location and that their value.

But only the difference is here, these weight or the importance is objectively calculated by using a model called Variogram. So, this lambda 0 here shows the vector of Kriging weights and Z is the vector of an observation at primary locations. And in a way Kriging can be also seen as a sophistication of the inverse distance interpolation.

(Refer Slide Time: 27:27)



So, what is the Variogram? Variogram is the model based on which these Kriging assign these weights or calculate these weights, how much weight is to use to give to any location which from the any location which is either close or distance from the unsampled location. So, this is what is being objectively calculated by variogram because in case of IDI the major problem was to determine how much importance should be given to each neighbor.

And it was arbitrarily decided by giving some values of beta however, in case of Variogram, the differences between the neighboring values are taken care of and at the same time, objectively we assign the importance using the Variogram. So, we generally calculate so, here you can see that he had the distance between the pairs of the points are given.

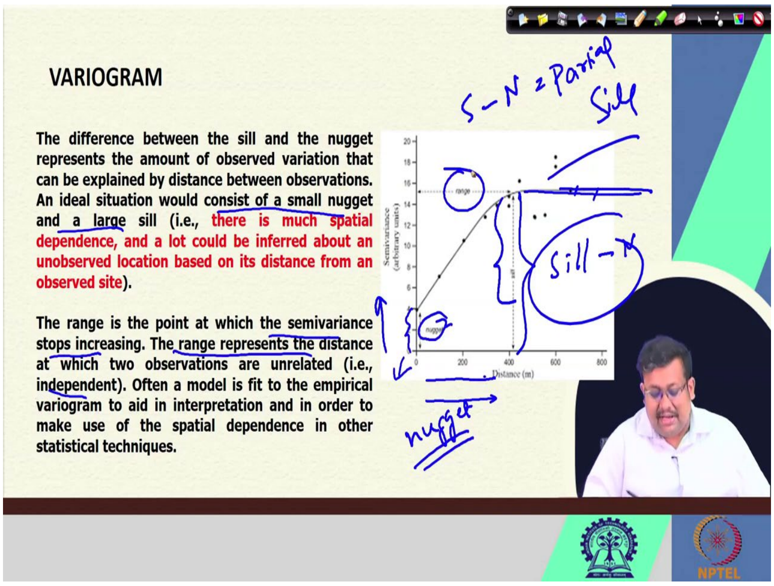
And in the y axis, we are putting the semi variance, what is semi variance semi variance is basically half one half of the square difference of the two of the properties between two sampling points. So, semi variance is basically measures of the spatial dependence between the observation at a function of the distance between them.

So, you can see with the change of the distance between the point pairs, we can see the semi variance are also increasing and it is say add after say after so, after a particular location it will reach a plateau. So, this is showing the spatial autocorrelation so, with increase with the distance up to a certain distance, these semi variance is related to the distance.

So, this is basically the graphical representation of spatial dependence or spatial autocorrelation after a certain distance there will be no further correlation it will reach a plateau and then there will be more or less constant. So, the semi variance is a measure of the spatial dependence between two observation as a function of the distance between them and the semi Variogram sometime we call it Variogram also.

So, semi Variogram is a graph which shows how semi variance changes as the distance between the observation changes. So, from this graph, we can see that how these semi variance will change with the distance between the observation pairs. Here the observation pairs means the pair considering unsampled location as well as the sample location.

(Refer Slide Time: 30:04)



So, in the Variogram there are different types of parameters, you should understand that here we are putting the distance and here we are putting the semi variance. Now, you can see here the difference between the sill and the nugget represents the amount of observed variation. So, here you can see at distance 0 there is some variability, which we cannot explain in terms of distance between the points.

This may be due to sampling error this may be due to measurement error and this is known as the nugget, n u g g e t, nugget component. So, nugget is the variability which shows the, mainly due to the measurement errors and the maximum variability is known as sill. And the point at which the semi Variogram reaches a plateau is the distance and this distance is known as a range parameter.

So, you can see the difference between the sill and the nugget represents the amount of observed variation that can be explained by distance between the observation. So, of course the sill versus minus nugget, if we subtract nugget from the sill this is the actual variation which we can observe which can be explained by the distance between the observation. So, these distance between sill minus nugget is known as partial sill.

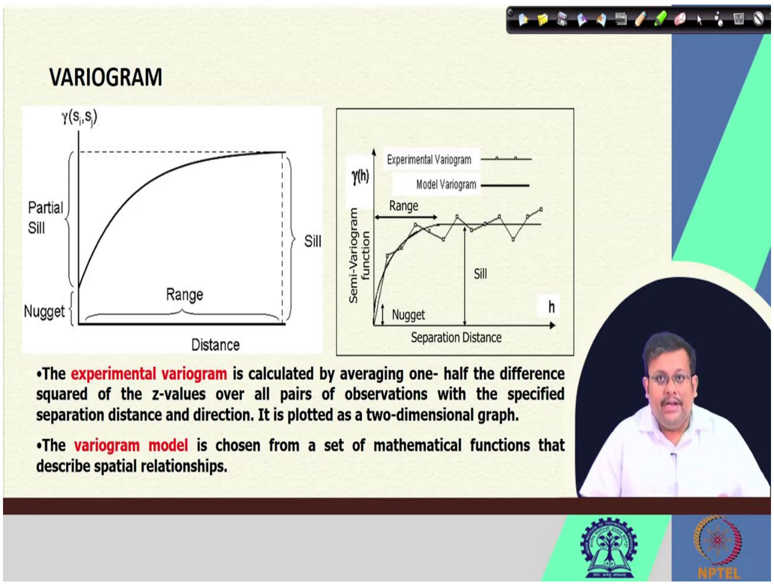
So, there is much special dependence. So, if course, an ideal situation would consist of a small nugget and large scale large scale. So, of course, we want that this partial seal should be maximized because and we should always try to reduce the size of the nugget so, that we can get maximum spatial dependence.

So, that means, there is much spatial dependence and a lot could be inferred about an unobserved location based on its distance from an observed site. So, this is what is ideal situation, but in some cases, you will see that the measurement error is too much that the most of this variation is accounted by this nugget.

So, that is in that condition, we will call it is a pure nugget effect. So, if nugget is showing the you know explaining the maximum of the scale or nugget is occupied a maximum of the sale then we call it is a pure nugget effect. So, the range is the point at which the semi variance stop increasing the range is represented by the distance at which two observations are unrelated that is independent you can see after this range, there is no relation between the distance and the semi variance.

So, here these points will be unrelated to each other after these range distance. So, often a model is fit to the empirical Variogram. So, basically we fit this model with some empirical Variogram to aid in interpretation and in order to make use of the special dependency in other statistical techniques.

(Refer Slide Time: 33:30)



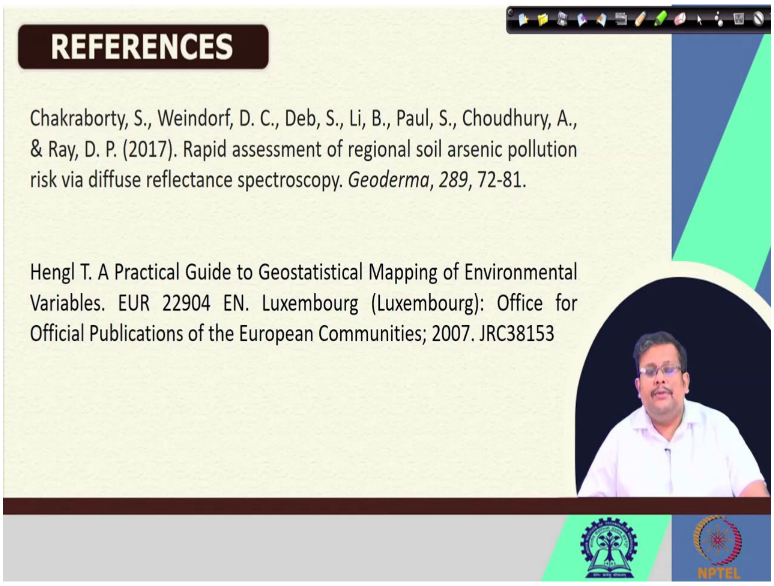
So, this is the Vario gram guys and we use the Variogram to model in the geo statistics. So, the experimental Variogram. So, experimental Vario gram is calculated by averaging one half of the distance squared of the Z values over all the pairs of observation with the specified separation this direction.

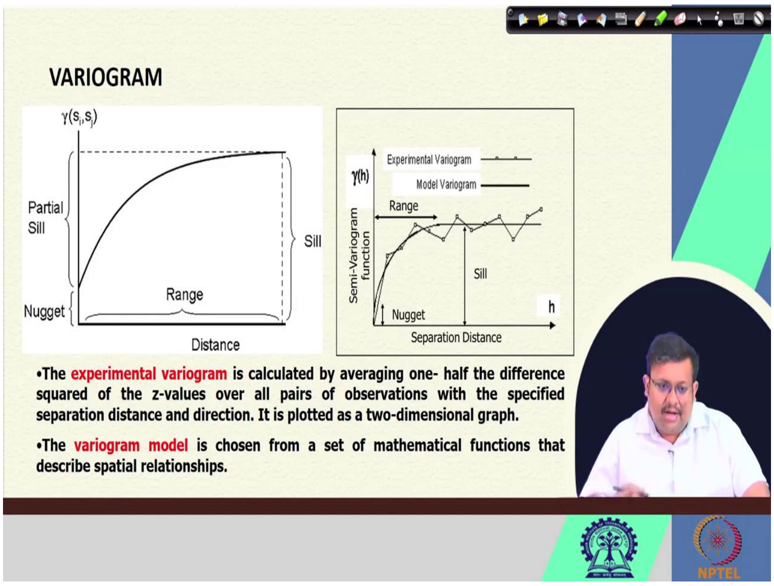
So, the distance if we just calculate the average of one half of the difference squared of the Z values that is property values of all pairs of observation, then we will get this semi

Variogram. So, we are putting the separation distance that means, the distance between the points in the y axis we are putting the semi variance.

So, semi variance means it is one half of the it is the average of one half of the difference squared of the Z values. So, we will get this type of ups and downs this is the experimental Variogram and then we try to fit these experimental Variogram using a model Variogram and this model could be of different types either exponential model or it could be different types of spherical model. There are different types of model.

(Refer Slide Time: 34:40)





So, guys, I hope that you have got some good knowledge about geo statistics of course, this is an overview, I would always strongly recommend you to go through some good literature for, geo statistics. Because geo statistics itself it is a discipline and it requires itself is a course. So, it is not possible to cover all these aspects in details within the short period of time.

But I would always recommend you to go through some of the literature to understand these geo statistical parameters and geo statistical modeling in a detailed manner and but I hope that whatever we need to understand for our for understanding the basic of DSM we have already covered.

And let us wrap up our lecture here, these are some of the references you can follow. Let us wrap up our lecture here and we will meet in our next lecture to discuss how to deal with a new software called R software and I will show you how to download and what is our and why it is extensively used in digital soil mapping, we are going to discuss thank you guys.