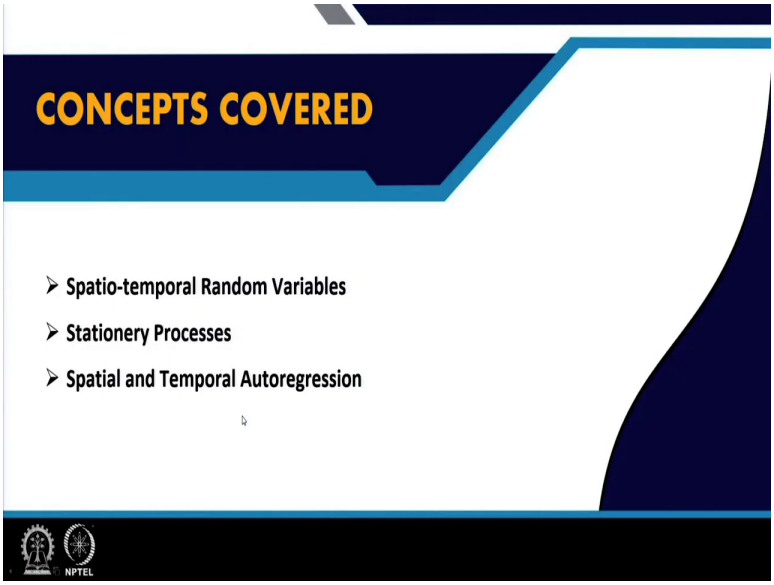
Machine Learning for Earth System Sciences

Prof. Adway Mitra
Department of Computer Science and Engineering
Center of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

Module - 01
Spatio - Temporal Statistics
Lecture - 02
Basics of Spatio-Temporal Modeling

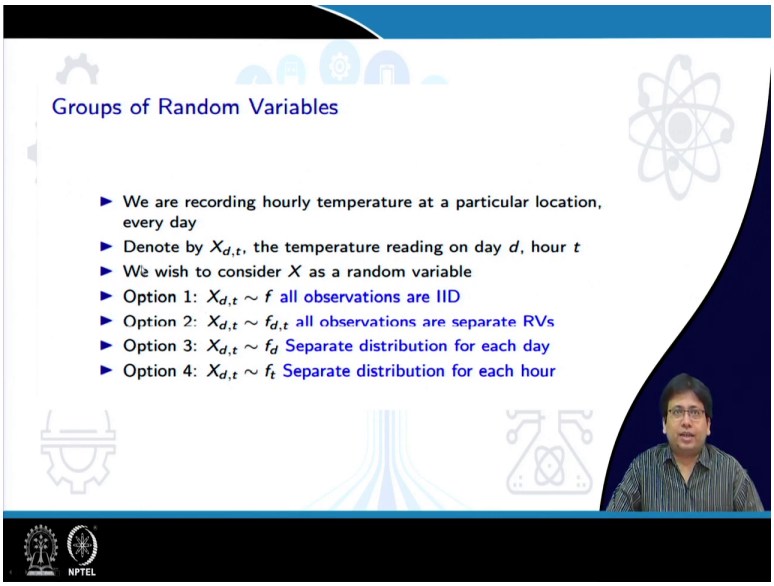
Hello everyone, welcome to the 2nd lecture of this module 1 of this course on Earth Systems Machine Learning for Earth System Science. So, in the first lecture we had mentioned that there are going to be 5 modules, the first module is going to be on Spatio-Temporal Statistics. So, in today's lecture we will go over the basics of this Basics of Spatio-Temporal Modeling and some basic concepts of this statistics.

(Refer Slide Time: 00:51)



So, the main concepts which we will cover today are first of all Spatio-temporal random variables which like we have already given a gentle introduction in the introductory video. And, then next we will consider the Spatio processes, and there the stationary processes. And, then we will go to Spatial and temporal autoregression concepts.

(Refer Slide Time: 01:13)



So, first of all like in the introductory lecture we already considered some template problems. So, we consider that let us say there is a particular geophysical variable which we are measure going to measure at different locations and different points of time. So, we like, let us say that we are recording hourly temperature at a particular location every day. So, just to remember that this, by particular location I may mean either a particular grid point or a particular like an in situ observation that is at a very at a single point.

So, like which one we are considering like is not important, it may like vary from one situation to another. So, let us denote the very these observations by $X_{d,t}$. So, by like the since the location is fixed. So, the this superscript S which denotes the location we can drop it because it is there is only one end it is fixed. So, just let us focus only on the time. So, we denote by $X_{d,t}$ where d means the day, and the t indicates the hour of the day.

So, its maybe day1 6 AM, day1 8 AM, day2 12 noon and stuff like that. And, so now, as we are mentioning this variable a or this observation $X_{d,t}$ we will imagine it as some kind of a random variable. So, either we can consider all the observations to be following the same distribution f. So, which basically means that these observations are IID; IID means Independent and Identically Distributed.

I independent means each of them is drawn separately or independently from a particular probability distribution; and identically distributed means like the, but that that particular probability distribution is same for all these like for all values of d and t.

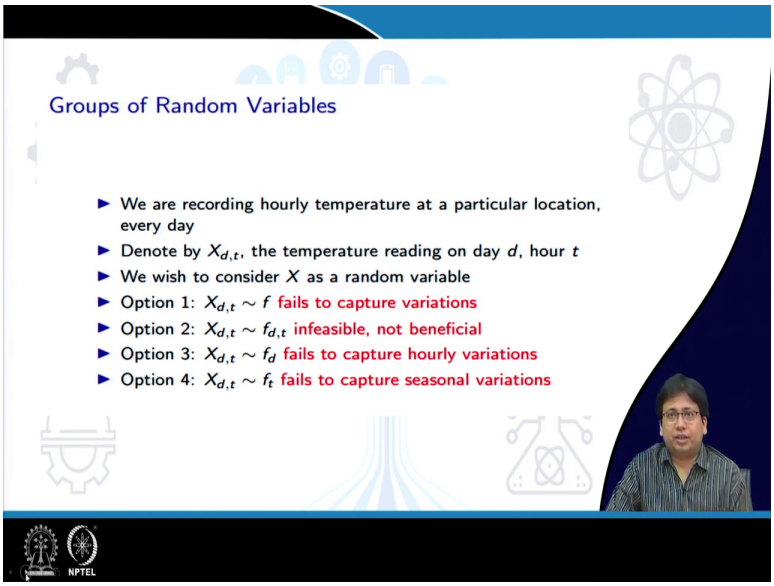
And on the that is one thing which we can do; the other extreme is just to consider that all the observations are separate random variables that is each of them follow a distribution of $f_{d,t}$ means that like they are independent, but not identically distributed. In fact, all of them are following different distributions.

The third possibility is we can consider a day specific distribution that is for any given day all the observations we have they for like follow the same distribution, but on a different day, it will be it may be a different distribution or at least it can be a distribute the same distribution may be with different parameters

And, X by $X_{d,t}$, the fourth possibility is that we have the separate distribution for every hour; that means, for everyday the 6 AM distributions will follow one distribution I mean the 6 AM observations will follow one distribution; the 8 AM observations will follow another distribution and things like that.

Now, when I say that they follow same distribution or when I say that they follow different distributions it may not be that one is a Gaussian and the other is a gamma or something like that. Maybe both are Gaussian or maybe both are gamma, but the parameters are different like that is enough for me to say that they are different distributions. So, either they can belong to different classes of distributions like one Gaussian and one gamma or they may belong to the same class of distribution, but the parameters might be different.

(Refer Slide Time: 05:05)



Now, each of these approaches they have their merits and demerits. So, if you consider the first one, the merit is rather obvious, you do not have too many distributions and you do not say you do not have too many parameters. Now, the common sense in any kind of modeling problem is that I do not want parameters; I do not want too many parameters. A few parameters is fine, but more the number of parameters the more complex is the model, I want to reduce the model complexity.

So, the first one has the least number of parameters, but at the same time it will also be able it will unable to capture the very different variations. From day to day there can be variations, from hour to hour there might be variations and things like that. So, if I consider that all the observations are following the same distribution, we will not be able to capture these kinds of variations.

Thus if we do the second one, it is like it is on one hand it is infeasible due to the extraordinarily large number of these parameters that will be needed; on the other hand it will not be beneficial also because if we consider that everything is independent, everything is separate not like anything else then there is no machine learning to be done. I mean machine learning is all about learning patterns and using those patterns to like I like to understand or make inference about those space, those locations and those times where we do not have observations.

So, if we the underlying assumption here is that, there is some kind of shared properties between different locations or different time points. So, if we are going for the option 2, we are basically saying that nothing is shared which is of course, useless. The third option like it where we are just considering one distribution for everyday, now this neglects the intra-day or the intra diurnal variations; that means, so suppose if you are considering the temperature on a given day, so, we know that like at midnight at 2 AM, 1 AM, 2 AM, 3 AM, the temperature will be very low. Then, as we move to the morning, say 6 AM, 7 AM, we will see a gradual increase in temperature. And, as we reach maybe 12 noon or 1, 1 PM the temperature might be maximum, after that maybe it will start dropping down.

So, this is something which we know almost always happens except some really unusual circumstances, but if we just think that the 1 AM temperature and the 1 PM temperature and the 6 PM temperature; all of them follow the same distribution then basically we are missing out on the on this hourly variations.

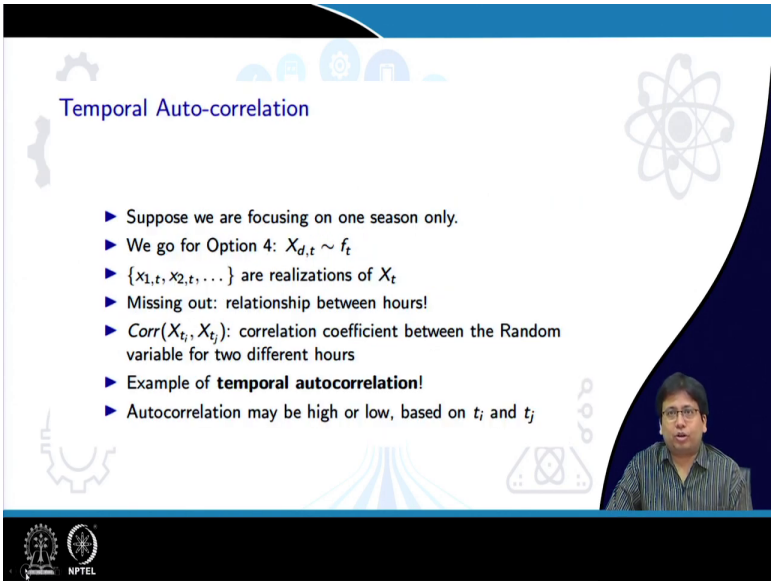
I mean you can of course, always the situation by adding a big variance on this that is like suppose you put the Gaussian distribution with a really large variance parameter so that all the values are feasible; however, that still does not stop or you stop this model from like making a forecast that the temperature at say 12 noon is much lower than the temperature at say 8 AM or some like in or unfeasible situation like that. So, this is not a great idea.

The fourth one, on the other hand, this can be an useful idea like here we are just saying that there is a particular distribution for 6 AM, there is another particular distribution for 7 AM, another distribution for 12 noon and things like that. And those distributions are same for all the days. Now, this can be useful only if we are considering the days within a particular season like say for the month of February if like we consider these kinds of this kind of situation that may make sense.

But thus like in general the temperature at 12 noon is should be very different in January and say in May or and in September. So, like there is in is like inter seasonal variations in this matter. So, this while it is able to capture the intra diurnal variations it fails to capture the inter seasonal distributions. So, all there none of these four things is perfect. Each of them has its own flaws. The first two are more flawed; the third and fourth are less flawed, but. So, like we may use third

and fourth only in certain say particular situations, but like both are also certain some kind of simplifications and they are not full proof.

(Refer Slide Time: 10:11)



Now, if you consider the one particular property of this like let us say we are considering the 4th option that is we are going for time specific distributions; that is we will say that the there is a distribution for 6 AM, there is another distribution for 7 AM, there is a third distribution for 8 AM and things like that. So, we saw that that has certain merits and we are which and this can be useful if we are focusing on one season only.

But, even then they are might like what it might be missing out on something some something more I mean ok. So, like we are considering that the like the 6 AM temperature of different days they follow from one distribution, the 7 AM temperatures of another of any day comes from another distribution. And the parameters of those distributions might be chosen in or fit estimated in such a way that the it is most likely that the 6 AM temperature will be less than the 7 AM temperature.

The 7 AM temperature will be less than maybe the 11 AM temperature which maybe more than maybe the 5 AM, 5 PM temperature and things like that. But by, but if we are considering all the different hours as separate like random variables I mean or following separate distributions, one

thing that is missing out here is the relationships between the hour that is like it might be implicit as I said from the choice of the parameters of this our specific distributions.

But, if we like suppose we know some kind of a relation we may not be able to that is we know that usually from 6 AM to 7 AM there is a 4 degree increase or from 7 AM to 10 AM there is another let us say 3 degree increase or things like that; we will not be able to bring in that kind of knowledge in this in like in this framework per se. So, what we are basically missing out is the correlation between the different the hourly random variables. So, let us say that t_i and t_j are two hours maybe t_i means 6 AM and t_j means 9 AM.

So, if you consider all the 6 AM temperatures and all the 9 AM temperatures there might be a correlation between them, that is, if I consider the correlation coefficient; and now I expect you to know what is correlation coefficient. So, it is possible that that correlation coefficient is let us say 0.7 or something like that is higher the 6 AM temperature is, we can expect the 9 AM temperature also to be correspondingly higher.

So, now, if we are going for this kind of a model where we are sampling the 6 AM temperature and the 9 AM temperature from two different distributions, we may not be able to have that kind of correlation between them, the kind of correlation which is observed in the data. So, this kind of correlation this is called as temporal autocorrelation. Now, autocorrelation this word autocorrelation we the word auto is used to mean that we are considering the correlation between the same type of variable.

It is like one both of them are temperature, both of them are the same geophysical variable X in general. It is not that one is temperature and the other is wind speed or something like that. So, if that was the case then I might have said cross correlation or simply correlation, but the auto here indicates that it is the same variation like same variable, we are considering the correlation of a variable with itself, but the only difference lies in time.

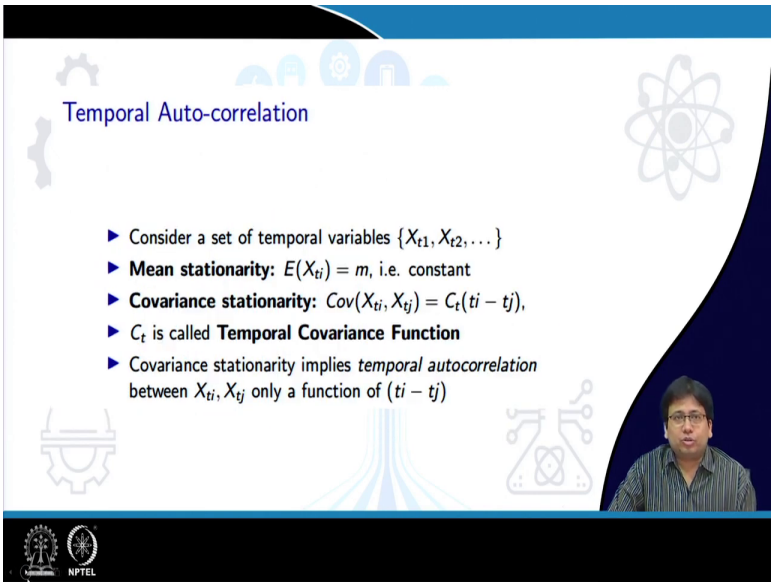
That is, we are measuring one variable at a particular time and the other let us say 6 AM and the other variable at I mean the same variable at a different time let us say 9 AM.

So, that is why this kind of this correlation coefficient between X_{t_i} and X_{t_j} I will call as temporal autocorrelation. Now, this autocorrelation may be high or low, it depending on t_i and t_j. Now by high like I when I say high what I mean is the magnitude not necessarily the sign that is while I will consider 0.9 as a high autocorrelation I will also consider minus 0.9 as a high auto correlation although its negative, but low means say something close to 0, ok.

So, now the when we are considering the auto this temporal autocorrelation of like of a physical variable like this like this autocorrelation may be either high or low depending on what these time points t_i and t_j are. So, in general if t_i and t_j are close to each other then we can expect high correlation because we can that is from our understanding of the process we know that if 6 AM is like the temperature is let us say 25 degree Celsius most likely the 7 AM will have a temperature of maybe 27 or 28 degree Celsius.

It is not going to be something very different like it cannot be that 6 AM is 25 degree Celsius and 7 AM becomes like 45 degree Celsius that is not possible

(Refer Slide Time: 15:40)



So, if t_i and t_j are close to each other then we can expect the other this autocorrelation to be quite high; now this brings us to the concept of stationarity. So, suppose I have the these two variables X_{ti} and X_{tj} that is the measurement of the same variable at temp at time t_i and time t_j.

Now, so, I can measure the covariance between them now suppose it or I can also at every time point I can also measure the expected value of them of these variables. Now, now suppose the expected value of this variable at all time points is the same, is a constant in that case I will say that the process follows mean stationarity ok. So, like. So, X_{ti} we have considered as a random variable. So, we had already considered that the time specific distributions are there. So, if the so, each of those distributions each of those probability distributions will have an expectation like, in case of the Gaussian distribution the expectation is same as the mean parameter

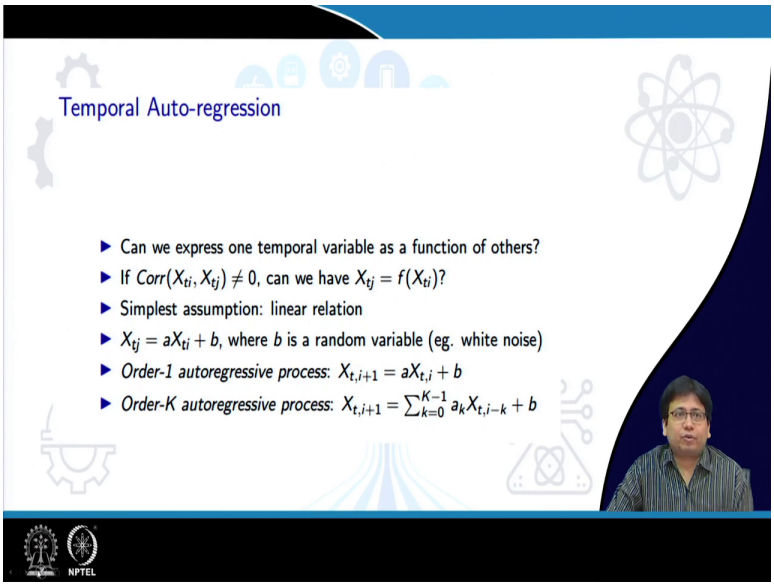
In case of gamma distribution or other distributions also there are some analytical expressions of this expectation thing. So, if all of these like our specific variables if all of them have equal expected value then we will say that the process, the probabilistic process defined by all these random variables it has the property of mean stationarity which means that the expected temperature at 6 AM is equal to the expected temperature at 7 AM is also equal to the expected temperature at 12 noon and so on.

So, in this case you as you can understand this is clearly not the case the temperature at 12 noon cannot be equal to the temperature at 6 AM. So, this in general in the example that we are considering it is not it does not follow mean stationarity. On the other hand, there is something called covariance stationarity. So, if the covariance between these two random variables is simply a function of the time difference between them then we will call it as temporal like we will say that it is like covariance stationarity with respect to time.

And this function, this C_t function this is called the Temporal Covariance Function. So, the so; that means, whenever two time points have equal difference between them the covariance between their corresponding variables is going to be equal that is the covariance between 6 AM temperature and 7 AM temperature is going to be the equal to the covariance between the 4 PM temperature and 5 PM temperature because both have equal difference of 1 hour between them ok.

So, in both cases the difference in what is 1 hour. So, what the that equal covariance is going to be that depends on a particular function C_t which is called the temporal covariance function. So, this also implies that the temporal autocorrelation which we were defined in the previous slide is also going to be a function of t_i and t_j ok

(Refer Slide Time: 19:04)



Now, based on the cons or hills of this concept of correlation comes the concept of regression. So, suppose like X_t they are covariance between two random variables X_{ti} and X_{tj} is non zero that is if that is the case then the question is can we then express one as a function of the other. So, that is can I find some kind of a function f such that $X_{tj} = f(X_{ti})$. Now, what kind of function then can that be? It can be anything, but so let us start with the simplest situation that it is a linear relation.

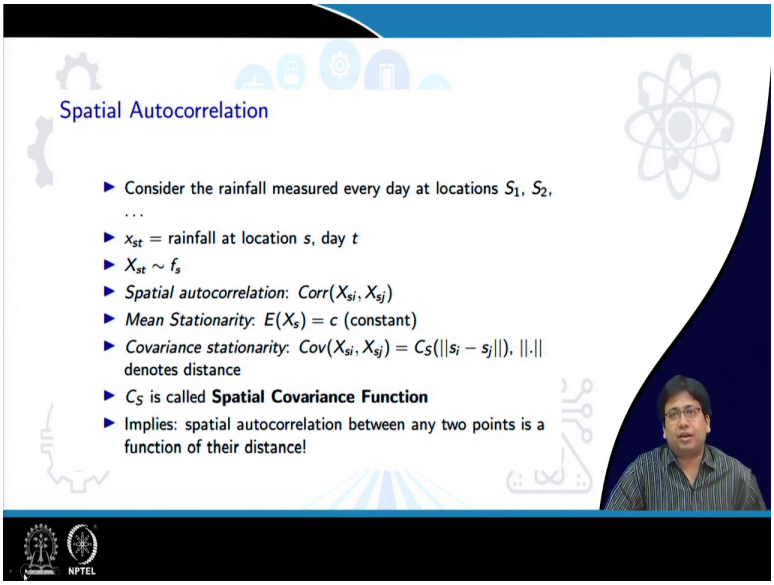
So, we say that $X_{tj} = aX_{ti} + b$, where a and b are some we can say that a is some kind of a scaling constant and b is a random is some kind of a random noise. So, it is like saying that the 7 AM temperature may be like 1.1 times the 6 AM temperature plus a small noise quantity similarly the 5 PM temperature may be 0.9 of the 4 PM temperature plus some noise quantity and so on.

So, that is now this is called an auto regressive process where like if the at every time point I am trying to predict the corresponding random variable or trying to express the corresponding random variable as a function of the previous time point. So, for 6 AM I will consider the 5 AM, for say 11 AM I will be considering as 10 AM and so on with a small like noise term.

On the other hand, I can generalize it to the k^{th} order auto regressive process in which instead of considering only the previous time point I will consider the previous k time points that is if I let us say if it is order 3 auto regressive process then for the say 8 AM temperature, I will be considering the 7 AM, 6 AM and 5 AM temperatures. And, and so needless to say as we increase this order of the auto regressive process the number of these parameters also keeps on increasing that is for each, so, these variables X_{ti} or X_{ti-k} these we can call as the predictors and each of the predictor has its own coefficient.

So, the more we increase the number of predictors the more there will be the coefficients; and it will be necessary for us to estimate all of these parameters from the data.

(Refer Slide Time: 21:44)



Now, similarly this the same idea we can also extend to the spatial domain that is instead of defining the covariance or the correlation between two observations at different time points we

can also consider we can also hold the time as fixed and consider the correlation between the observations at different locations let us say s_i and s_j.

So, like so, let us say that we are considering X the like X_{st} means the rainfall at location s and day t. So, like in the introductory lecture, I had put this s the location as the superscript, but like for convenience let us just consider both the location and the time as subscript. So, the first one in this case indicates the location s and the second one indicates the time t. So, now, let us say that these X_{st} it is not specific to some location.

So, that is the it is independent of time, but it is dependent on location. So, that is why I have written as f_s that is every location has a corresponding distribution over the over the variable.

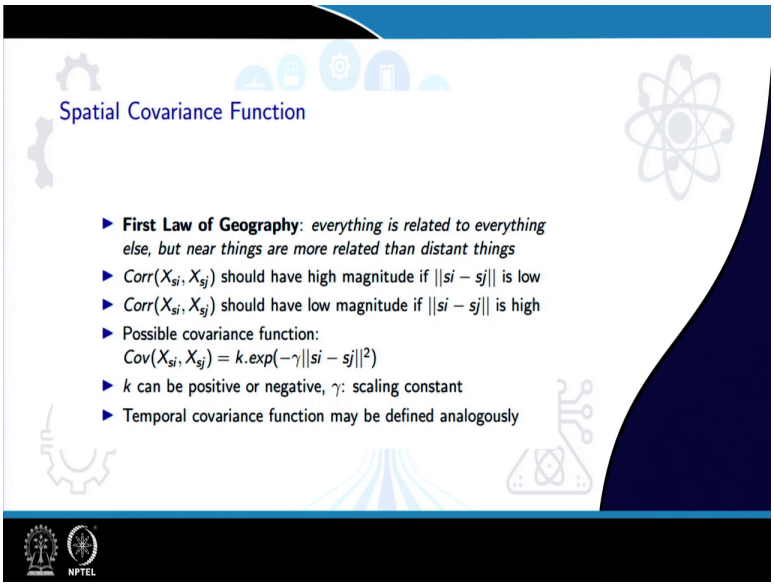
Now, we consider the what we what we can consider is the is the correlation between the X at different locations. So, let us say you have s_i and s_j two locations and we are considering the covariance or the correlation between the corresponding variables.

So, again we can define the concepts of stationarity. So, mean stationarity will basically mean that the expected value is same at all locations. It is independent of the locations and covariance stationarity will mean that again just like the temporal covariance stationarity; if we consider any to the variable or the observation at any two locations; and if we consider their covariance that covariance is going to be just a function of the geographical distance between the two those two locations.

So, in the previous case we had the difference between those two time points; in this case we are having the difference that is to say the geographical distance between the two locations. So, any two locations which are equidistant can expect to have roughly the same covariance and it is yeah. So, this C_s this in this case is the called the spatial covariance function. In the previous case we had a temporal covariance function in this case we have a spatial covariance function.

So, this basically means that the spatial autocorrelation between any two points is a function of their distance.

(Refer Slide Time: 24:38)



And, like we may in general we may expect that if two locations are close to each other then they may have a very high correlation; and if they are far away from each other they may show very low correlation. Again, by high or low I am talking about the magnitude. And so like this is sometimes known referred to as the first law of geography which says that everything is related to everything else, but near things are more related than distant things.

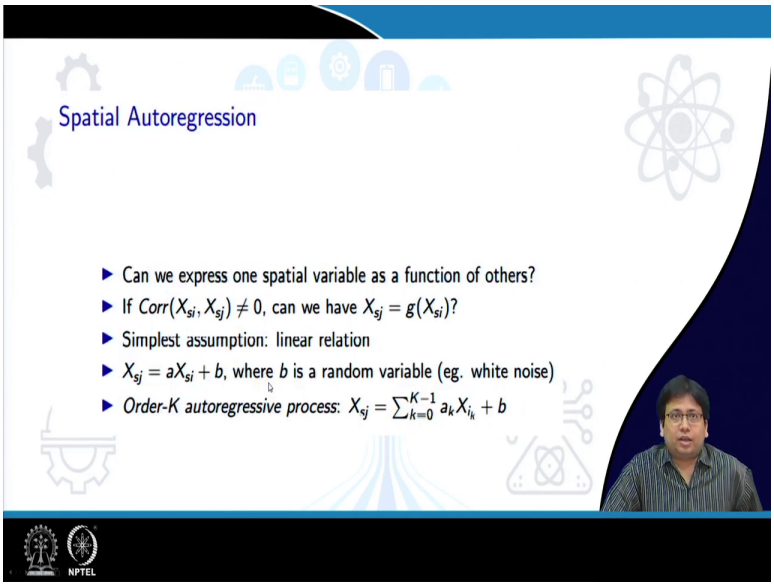
So, like suppose I am considering the temperature between the temperature at Kharagpur and in Kolkata and I am measuring the covariance between them. So, I can expect that that covariance to be quite high because they are very much dependent on each other. It is when we know that whenever Kolkata has high temperature most likely Kharagpur will also have high temperature and so on.

So, we can if you consider their correlation coefficient, it may be something close to 0.8 or 0.9. So, some of you may actually do this as an exercise and actually grab some data and take different pairs of locations which are close to each other let us say Kolkata, Kharagpur or maybe Chennai, Bangalore or Mumbai, Pune and things like that and see whether what their that correlation between their various geography geophysical variables are.

On the other hand, if we are considering let us say Kharagpur and London or Kharagpur and New York, we can expect a very low correlation. I mean by low correlation I do not mean say minus 0.9 or something like that I mean something like 0 ok. So, in general when is that this distance is low, I can expect high correlation and when there is high we can expect low correlation. So, one possible example of this covariance function can be this one like which satisfies this kind of a property.

So, the temporal covariance function may also be defined like analogously. So, like when we because which follows the same properties.

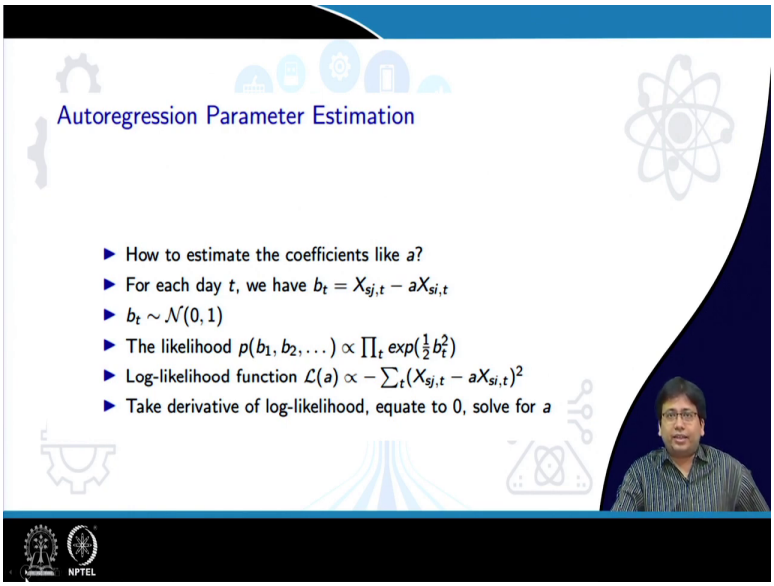
(Refer Slide Time: 26:41)



Now, just like we considered the temporal autoregression similarly we can consider the spatial auto regression also where we may want to express the observations at one location in terms of the observations at some other locations that is I can like express $X_{sj} = g(X_{si})$. And, once again just like the previous case the simplest assumption is the linear relation like something like this where a and b like where a is like a scaling constant and b is some kind of a random noise ok.

And in this case also we can instead of an or like a simple order one auto regressive process we can go for a order k auto regressive process where we are using the k location the measurements at k different locations as predictors for one particular location.

(Refer Slide Time: 27:38)



And so, then the question arises how to estimate the coefficients like a? So, the as you know when we have linear regression problems we can estimate their various parameters using the least squares method. We can also look at that method in a slightly different way. So, by focusing on these random quantities the random noise.

So, at every time point or at location every location we have this noise parameter which is the difference between the its actual value, the observation and the and its expected value which is the like this aX_{si} and we know that this noise they follow the Gaussian distribution. So, let us say that we put this we imagine they follow the Gaussian distribution like this.

So, we what we do is we try to like use the maximum likelihood function. So, we try to choose the these parameters in such a way this parameter a in such a way that the joint distribution of all the daily noise is maximized. So, we write the joint distribution. So, remember that they are all in these noises they are all independent. So, their joint distribution is nothing but the product of their individual distributions which is which is just this Gaussian distribution and here we b_t we know that b_t can be replaced can be expressed in this way.

So, I can like estimate the parameter by solving this maximum by we can just write the derivative equate it to 0 and solve for the parameter a.

(Refer Slide Time: 29:21)

And yeah so, for a good reference for all the materials which we covered today is this handbook of special statistics by Gelfand and all you will find all these necessary concepts with various examples. So, that brings us to the end of this lecture.

Thank you everyone. We will continue our discussion on Spatio-temporal statistics in the subsequent lectures. Bye, bye.