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

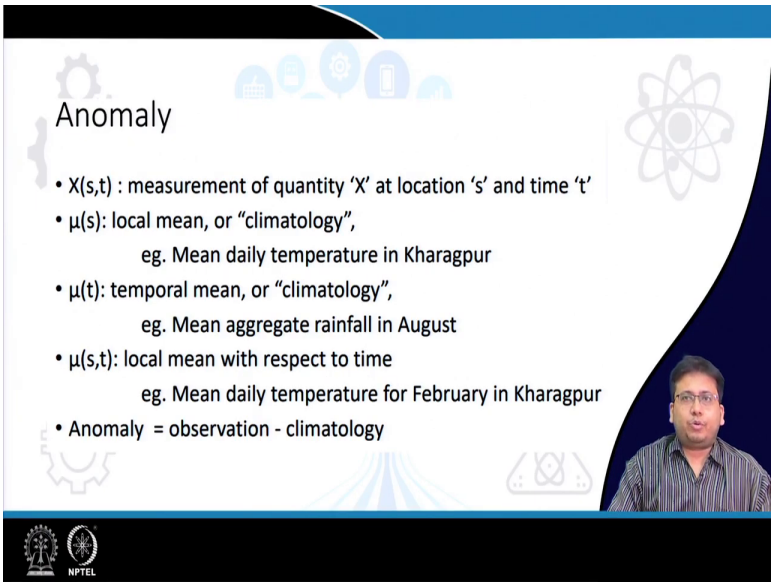
Module - 01
Spatio - Temporal Statistics
Lecture - 05
Anomaly Event Detection

Hello everyone. Welcome to lecture 5 of this course on Machine Learning for Earth System Science. We are currently in module 1, which is about Spatio-Temporal Statistics and the topic of today's lecture is Anomaly Event Detection.

(Refer Slide Time: 00:40)

If the concepts we will cover in today's lecture are anomaly and anomaly events, clustering based anomaly detection and probabilistic anomaly event detection.

(Refer Slide Time: 00:52)



So, first of all let us try to understand what the this word anomaly means. So, anomaly basically means something unusual, some deviation from what is normal. Now, in the specific context of earth sciences, anomaly is anomaly has a particular a definition, so let us say that $X(s,t)$ this is the quantity of our interest, we like we are measuring a variable X at location s and time t.

Now, this let us denote by $\mu(s)$ the local mean or the local climatology, like which is a specific for that particular location for example, the mean daily temperature in Kharagpur. So, here the locate this is specific to a location, say Kharagpur, but it is not specific to time, it is the mean that we are talking about has been taken across time.

Similarly, we can also define the temporal mean which we denote by $\mu(t)$, it can also be called as the temporal climatology example, the mean aggregate rainfall in August ok. So, again see that here this is not specific to any location, but it is specific to some time, that is the month of August. Now, $\mu(s,t)$ basically means the local mean with respect to time that is the for example, the mean daily temperature for February in Kharagpur ok.

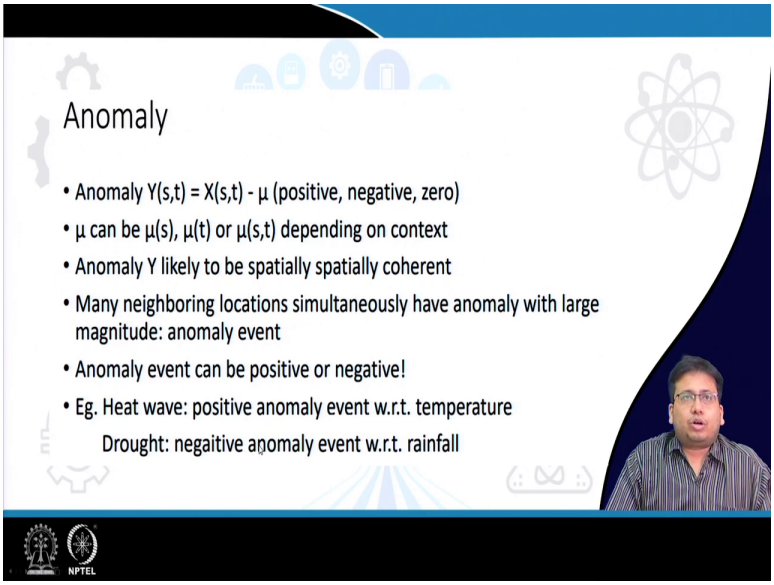
So, we can define these kind of climatology as the long term behavior, now that can be specific to location, specific to time or specific to both location and time. Note that when we are talking about the temporal mean that the time we are talking about like has to be something like a month

which recurs repeatedly. It cannot be something like we cannot say a temporal mean for a year, because a year is like is happens only once. I mean one a like it does not recur, but month is something that recurs. So, it makes sense to define the temporal mean in for months, but not for years ok.

So, now what is anomaly? So, anomaly is basically the observation minus the climatology.

$$Anomaly = observation - climatology$$

(Refer Slide Time: 03:12)



So, for example, let so let us consider the $X(s,t)$ the observation of variable X at location s and time t. So, anomaly that in the corresponding anomaly which we denote by $Y(s,t)$.

$$Y(s,t) = X(s,t) - \mu.$$

Now, like already we have defined three kinds of μ or climatology: spatial, temporal or spatio temporal.

So, here which μ we are talking about will depend on the context, it can be any of these. So, needless to say this anomaly Y it can be positive, it can negative or even 0. Now, 0 means the a the $X(s,t)$ is perfectly equal to climatology which is of course, very unlikely or next to

impossible, if these are real numbers. If these are discrete quantities then it is possible, but usually like in earth science we the measurements are not discrete, they are continuous.

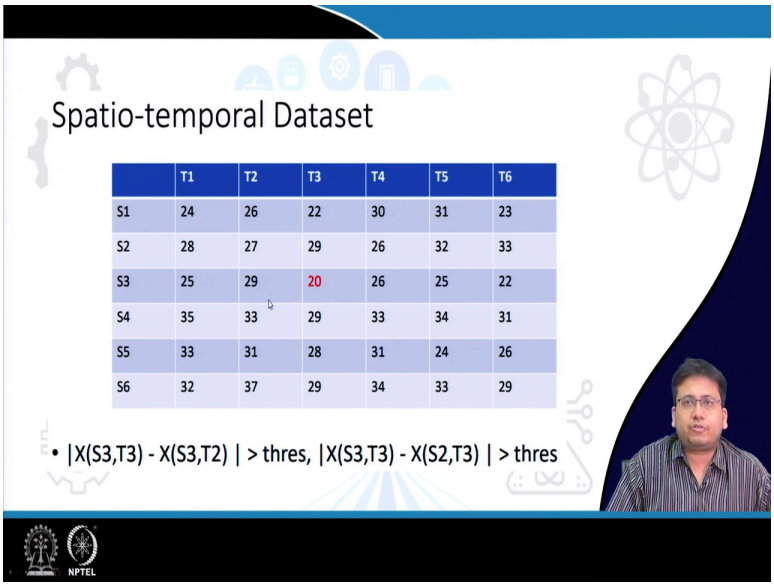
Now, these so when like an and one particular location is having some kind of anomaly at a particular point of time, let us say a positive anomaly. So, we can say that if the temperature today is let us say $32^0\ Celsius$, while the mean temperature at Kharagpur in the month of March is let us say $30^0\ Celsi$us; that means, that we are currently having a $+\ 2^0\ Celsius$ anomaly.

Now, it happens that these anomalies are usually spatially coherent; that means, that like if one location is having some value of anomaly, neighboring locations are also likely to have similar values of anomaly. So, if the temperature anomaly it at Kharagpur is let us say $2^0\ Celsius$, it is possible that Kolkata which is not far from it will also have a similar temperature anomaly, the same thing can be said in case of with respect to time also.

Now, when many neighboring locations simultaneously have an anomaly of large magnitude, such a situation is called an anomaly event. For example, a heat wave, say suppose several locations nearby are having a like a like positive temperature anomaly, that can be called as some kind of heat wave of course, there are specific operational definitions of heat wave that is how much the anomaly should be and so on some thresholds are there.

But the general idea is that like an anomaly event is something when several nearby locations or like several neighboring time points have a high value of you know of an anomaly and that persists either for a for some time or it extends over space. And this anomaly event also can be either positive or negative. We have already considered the example of heat wave. Another example is a drought, which is basically a negative anomaly with respect with respect to precipitation.

(Refer Slide Time: 06:28)



So, now suppose we have a some kind of a spatio temporal data set. So, the task here which we are talking about is anomaly detection or anomaly event detection. The now, let us say we have a large volume of data which is like either observation or model simulation and we want to identify these kinds of anomaly events ok. So, first of all we will start off by identify like calculating or identifying individual anomalies. So, this is what our data set looks like.

So, there let us say there are 6 locations, $S1$, $S2$, $S3$, $S4$, $S5$, $S6$ and there are similarly 6 time points $T1$, $T2$, $T3$, $T4$, $T5$, $T6$. And let us also consider that the first three locations, $S1$, $S2$, $S3$, they are spatially close to each other the next set of locations $S4$, $S5$, $S6$, they are also like another set of spatially closed locations. And here let us say we have observations of a particular variable which may be temperature or anything else also at these locations and this time.

So, the question is like I want to identify various anomaly events in this kind of a spatio temporal data set. So, how do we go about it? So, as a like remember the definition; the definition is that like the $Anomaly = observation - climatology$. Now, what is the climatology we need to consider in this case?

So, like here one possibility is at every location we like calculate the temporal climatology by taking the like or we calculate the spatial or for any location we, either we calculate the spatial climatology by taking the average over all the time points or on a particular time we take the we or we calculate the this temporal climatology by taking a mean over all the locations. These things will make sense only if our data set is large enough.

Because when we talk about climatology, we are talking about some kind of mean over or like we are talking about some long-term mean. So, either it has to be over a large number of locations or it has to be over like over a long time period ok. But in case we do not have so much of data we cannot calculate the climatology directly from the from that data set. So, instead we have to do some other things such as the we have to look for localized that is we have to look for localized clues of anomalies.

So, for example, like we it might we may remember that these anomalies they tend to be or like I mean not anomalies I mean any kind of geophysical variable tends to have a certain degree of spatial and temporal correlation or coherence whatever you may say. That means, that it is unlikely that a particular value is going to be very different from its spatial and temporal neighbors.

So, one way of identifying anomalies in the absence of these climatology values may be to simply compare every measurement with the measurements of its spatial and temporal neighbors and whenever we see a large deviation we can call it as a an anomaly. Say for example, this one at location $S3$ and time $T3$. So, here the observation a measurement is 20, which is significantly different from that of like, on time $T3$ its neighboring locations $S1$ and $S2$ they have values like 22 and 29.

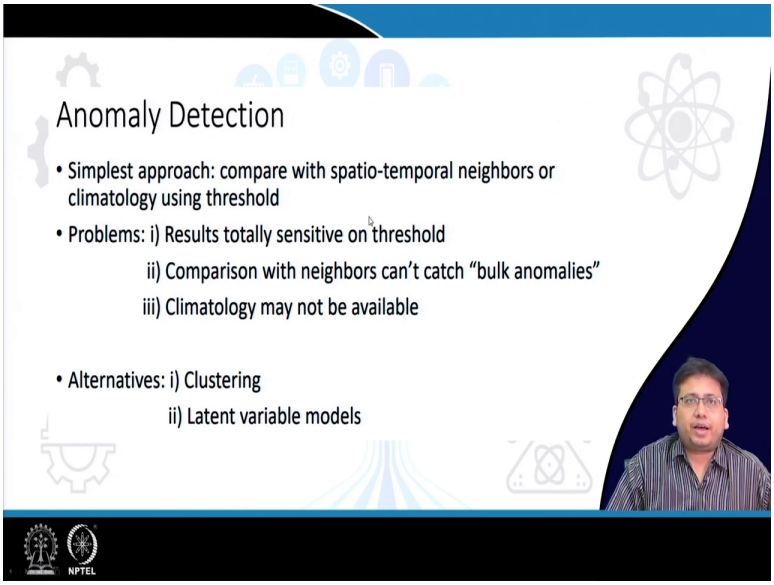
While this one like if we consider S for the same location $S3$, if you consider other time points say $T2$ and $T4$ here also we see big differences. So, we may tend to conclude that this is an anomaly. Now, this is not a subject I mean this is not a very objective there are certain subjectivities involved here, say, for example, the difference between 29 and 20 here is 9, but the difference between 22 and 20 here is only 2.

On the other hand, like if you consider on 29, here its difference from this 20 is 9 and its difference from this 22 is 7. So, why should we not consider this 29 as the anomaly rather than this 20 as the anomaly right. The one answer might be that if you consider the like the temporal differences at that same location $S2$, there we see that 29 is actually not very different from 27 and 26.

But if you consider location $S3$ then 20 is quite different from both 29 and 26. So, that way like we may tend to think that it is this which is the anomaly, but not this 29 ok, but again there is some kind of subjectivity involved here. So, we are basically comparing again the differences against thresholds, but what is the value of threshold; obviously, whether something is being called as an anomaly or not will depend on this definition of threshold.

So, how to choose that threshold? So, there is no universal answer to this question. Then similarly, we can also like in the same data set we can also identify some other anomalies by comparing the neighborhood differences with these kinds of thresholds and we can find more and more of such anomalies.

(Refer Slide Time: 12:31)

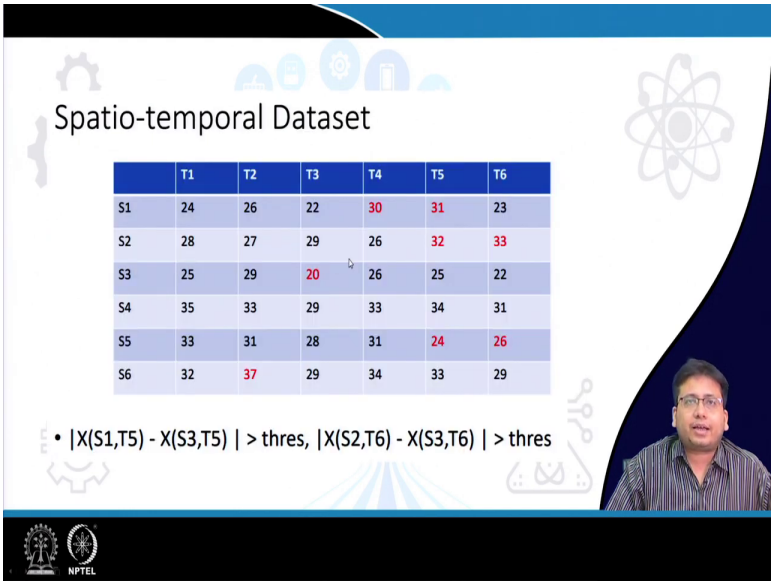


So, basically this is the simplest approach, we compare every value with the with its spatio temporal neighbors or the climatology if that is available using some kind of threshold.

Now, the problem here is that the results are totally sensitive on threshold, like if some if the threshold value is changed let us say from 5 to 6, then several of the anomalies will disappear. Again, if I slightly lower the threshold say from 5 to 4, then lots of new anomalies may come up, so that is one problem. The second is that the comparison with neighbors cannot catch the bulk anomalies.

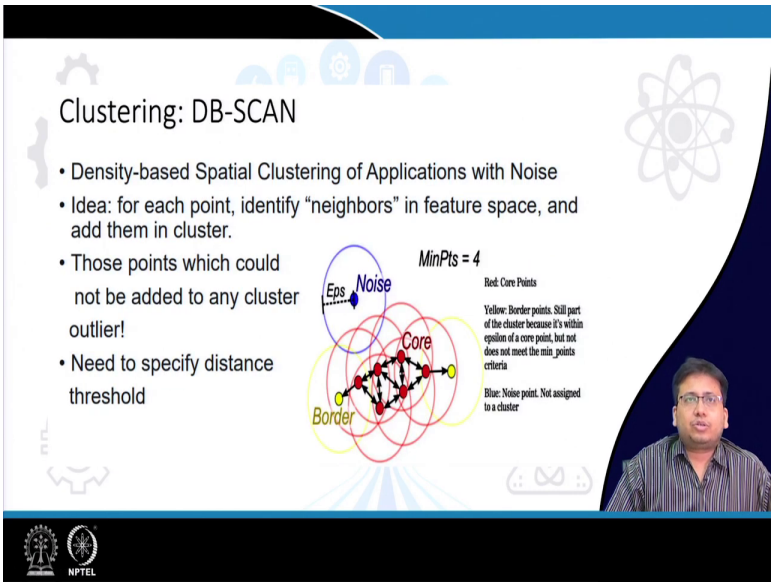
So, basically an anomaly event I mentioned an anomaly event is a situation where several locations are simultaneously affected, they are simultaneously having like deviation like their values are simultaneously deviating from their respective climatologies. However, if we like because the neighboring locations are all affected together, if we take the differences with the neighbors we may not be able to catch that such a thing has happened. So, this threshold based method will not work in that case.

(Refer Slide Time: 13:47)



And finally, of course, the climatology may not even be known like a like for example, in this data set, we did not know what is the climatology. If we knew then we could have of course, use the definition, but we may not know it if we do not have enough data ok. So, there alternatives are either to use clustering or to use latent variable models, both are ideas borrowed from machine learning.

(Refer Slide Time: 14:08)



So, typically for clustering now there are lots of clustering algorithms in machine learning. So, k-means clustering you have all heard of, but this is not suitable for k-means clustering, because here we are it is not that we are trying to create a set of homogeneous clusters, rather we are trying to like identify those values which do not lie any in any cluster. It is like we are trying to identify the values that are basically something like the outliers.

Now, k-means is not the ideal way of identifying outliers for various limitations related to its initial conditions and things like that. Now, for this purpose of identifying these outliers or a like anomalies in the like in case of spatio temporal data, one algorithm which is used frequently is the DB-SCAN algorithm this is the Density-Based Spatial Clustering of Applications with Noise, that is the full form of DB-SCAN.

Now, one obvious example, an obvious advantage this DB-SCAN has over k-means is that, in k-means algorithm it is required to specify k that is the number of clusters being formed and for and in most situations we do not have the slightest idea of what k can be. And this problem is going to be further like further amplified in case there are anomalies or outliers in the data set, which is precisely the case here.

So, what we go is some kind of this non parametric approach where we do not specify k or anything like that initially, but we actually try to grow the clusters. So, the idea is as follows, for every point you identify its "neighbors" in the feature space. So, the feature space in this case may mean the location, time as well as the values of the observations and any covariates if there are. So, now whenever we find these kinds of "neighbors" we add them into the cluster.

Now, those points which could not be added to any of the clusters will be considered as an outlier or anomaly. So, the whole things proceeds like this. So, let us try to understand this with some kind of with an example here. So, let us say that we have an initial point, like we search for neighbors other points which like which are close enough to it in the feature space. So, we lets say we find this point and which is close enough. So, I just join these two.

Next, I will search in the vicinity of this original point as well as the new point and let us say I find this yellow point, this point and which is close enough to this and along with it I also find this one. So, I add both of them to the to my cluster which is growing, then further I search in the vicinity of this point, but I cannot find any more point. So, I just stop here. Now, I search in the vicinity of this point and I find let us say these two new points.

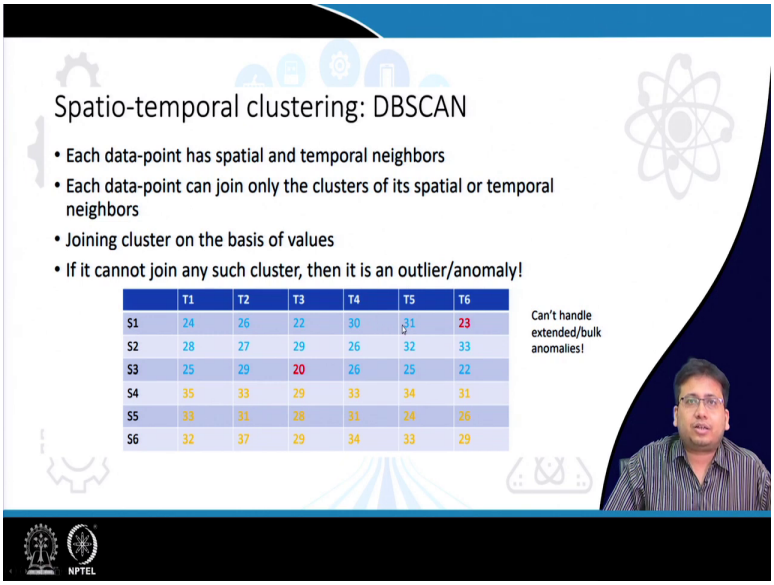
So, again I add them to the cluster and I search in their respective vicinities. Again, from this point I find this new point and I add that, but from when I search in the vicinity of this point I cannot add any more so I stop. So, basically whenever there is a new point to which I cannot find any more neighbors who which can be added to the cluster then like we call those points as the boundary points, while the points which could add more neighbors to the cluster they are called the core points.

So, finally, the like we started with an initial point and we like progressively added more and more points in the cluster. So, now, suppose one cluster is completed, I cannot grow it any further. So, some more points will be left here and there. So, I will next focus on another point and again I will repeat the same process, I will search in its neighborhood and see if there are some neighbor, neighboring points which I can add to the cluster and similarly that cluster I will grow as long as I can.

Once I have finished growing that cluster, then I will focus on the remaining points which have not been added to either cluster and try to grow clusters around them and so on. I will repeat this process, till I am left with only such points which cannot be added to any of these clusters like this. So, such points we can call as the outliers or the anomalies. So, this is the DB-SCAN algorithm, but one problem it has is rather obvious: we have to specify the distance threshold for any point, two points to be added as neighbors.

So, in a sense this is not very like this the same problem which we faced in this case remains in the DB-SCAN case also. Of course, DB-SCAN may have more advantage have certain advantages compared to the like this kind of an approach, but essentially the idea remains the same, but and both of them have the same problem namely we have to specify the thresholds.

(Refer Slide Time: 19:49)



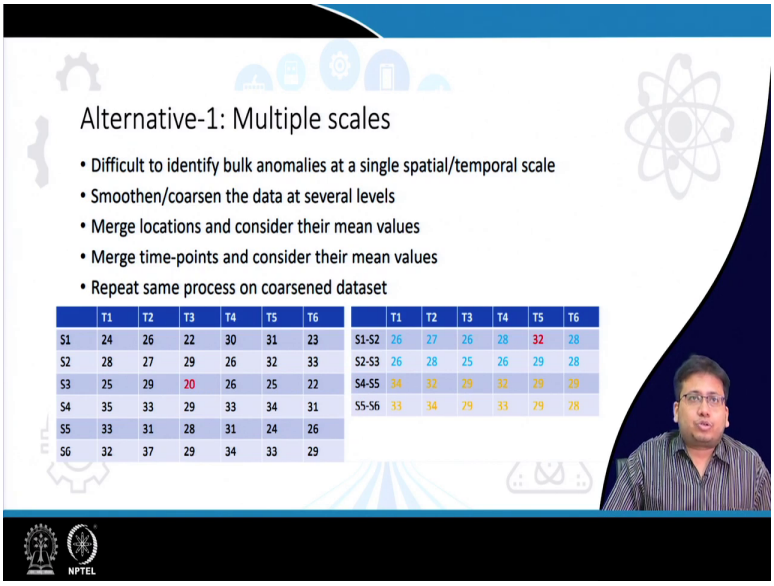
There another problem is that, it may not we may not be able to find like these bulk anomalies or anomaly events so well using the DB-SCAN algorithm. That is like suppose like a large number of like, here if we can see some kind of there is some kind of an anomaly event which covers or which is affecting these locations. But so like here as you can see like at location $S1$, initially the values are 24, 26, 22 then suddenly it jumps up to 30 and 31 after that it again falls to 23. Similarly, in $S2$ you can see that here in $T5$ and $T6$ there is a sudden jump.

Similarly in S in S in case of $S3$; however, this kind of a thing is not seen. So, you can say that there is a bulk anomaly which or an anomaly event which impacts $S1$ on $T4$ and $T5$ and $S2$ on $T5$and $T6$. So, we can say that these four things are part of the same anomaly event or like these four define an a particular annual event. The problem is that because of the nature of the like, so, note that like it is this is the anomaly event is not localized in either space or time, that is it starts in $S1$, but it affects later effects $S2$.

Similarly, it starts in $T4$, but it splits over to $T5$ and $T6$. Now, because of this reason if you run the DB-SCAN algorithm on this data set you will find that it is unable to capture this bulk anomaly. That is these values actually get absorbed within a larger cluster like this ok, this like the these all these points, these blue points. These are like you can say that these are one cluster and these orange points, these are another cluster a red and this red and this red are identified as some kind of anomaly.

So, the funny thing here is that this 23, this actually may not have been an actual anomaly because at location $S1$ this it is like it might be that this 23 is actually it is it is close to its mean very mean value; however, because in case of DB-SCAN since we are considering only or comparing only with the neighbors here because this is following an anomaly it will seem that this is the anomaly rather than this. So, this is the problem of localized approaches like this.

(Refer Slide Time: 22:34)



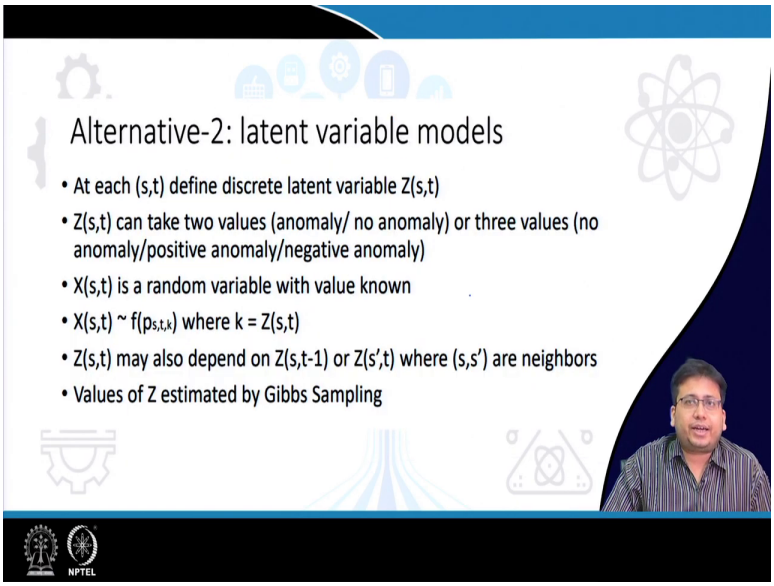
So, now what is the alternative to this problem? One alternative is to do the anomaly detection in at multiple scales. So, I one possibility is that we coarsen or smoothen the data at like several levels, either spatially or temporally or both. So, here like what we have done is that we have merged locations, that is instead of considering $S1$ all the 6 location separately, like I consider the like I merge $S1$ and $S2$.

Similarly, I merge $S2$ and $S3$, I merge $S4$ and $S5$ etcetera, etcetera, etcetera and then on this merge data set I then do the anomaly try to do the anomaly detection and see if that helps. Now, in certain situations that might help, but like in this particular case as you like as you can see, it helps only to a certain extent that yes I do find that there is an anomaly here that is $S1$, $S2$ is having an anomaly on $T5$, but we are unable to understand that same anomaly event actually starts in $T4$ and ends in $T6$, starts in $S1$ and ends in $S2$ that nuance is lost.

We understand that something unusual has happened there, but we are unable to like get it, on I mean get its full extent. On the other hand, certain isolated anomalies or which have lesser spread they might get like smoothened out in this process ok. So, like it is necessary, so if we are going for this kind of an approach it is necessary to do it at multiple levels, that is at different like at this level of course, smoothness we identify a set of anomalies.

Then we smoothen the data a bit further and identify another set of anomalies, then we further smoothen it and get another set of anomalies and so on. And finally, we find a way to link the anomalies found at different levels. So, that is one way of solving this, anomaly detection problem, I mean anomaly event detection problem.

(Refer Slide Time: 24:51)



The other approach is through latent variables. So, you will remember that while discussing the spatio temporal stochastic models, we had introduced this concept of latent variable.

So, we can do the same thing at every location or like at every point (s, t) we can define discrete latent variables $Z(s, t)$. So, this $Z(s, t)$ its a discrete value which can take either, I mean it is a discrete variable which can take either 2 values or 3 values. 2 values as in anomaly or no anomaly and three values as in no anomaly, positive anomaly and negative anomaly. So, depending on the observations, we will try to like estimate the values of Z and from that values of Z we will understand where the different anomalies are located.

Now, this $X(s, t)$ we can once again consider it as some kind of a random variable with whose value is known, that is we know an instance of the instantiation of the random variable. So, this is the same idea as we considered in the previous lectures, where we are considering or we are imagining the observations as realizations of a random variable. So, what like in this imagination we can consider that $X(s, t)$ it is a random variable, it follows our distribution which is specific to the corresponding value of Z.

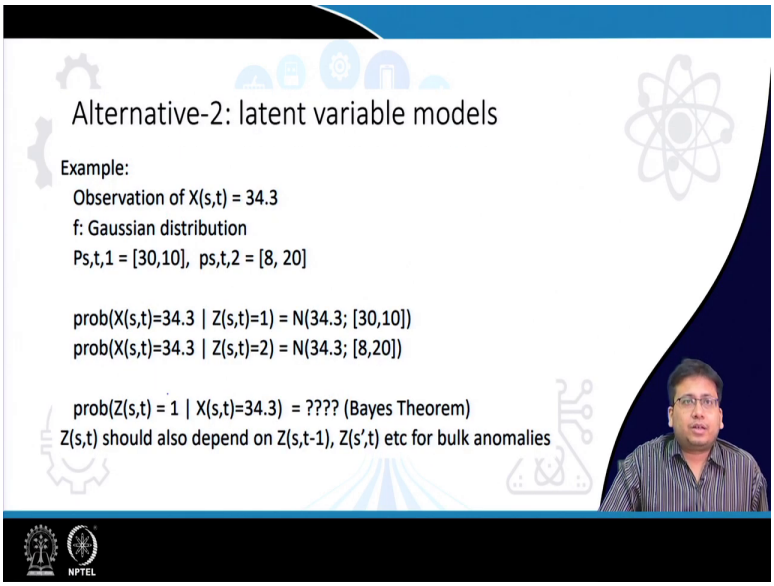
That is whether Z is an anomaly or not or, if whether Z is a positive anomaly, negative anomaly or no anomaly. So, in each of the situations, like we will consider that $X(s, t)$ has a will follow a

particular probability distribution. Let us say for like positive anomalies $X(s,t)$ may have a certain range of values. For negative anomalies it may have a different range of values, for no anomalies it may have like a another range of values which lies in between the two ok.

So, the task here is that given the observations of $X(s,t)$ and a given or like with certain parameters or thing and things like that which may have either been set by the modeler or estimated somehow from the data. We will solve the inverse problem that is given X, we will try to estimate the values of Z through approaches such as Gibbs sampling.

And the another thing which can be achieved in this case is that, we can make the this or we can define this random variable $Z(s,t)$ in such a way that it also depends on the spatial or temporal neighbors of this of this observation. So, $Z(s,t)$ may depend on the $Z(s,\ t-1)$ which is the situation at the same location at the previous time point or at $Z(s',t)$ which means the situation at the same time, but at neighboring locations. This is because we know that these anomaly events are spatially and temporally contiguous.

(Refer Slide Time: 28:08)

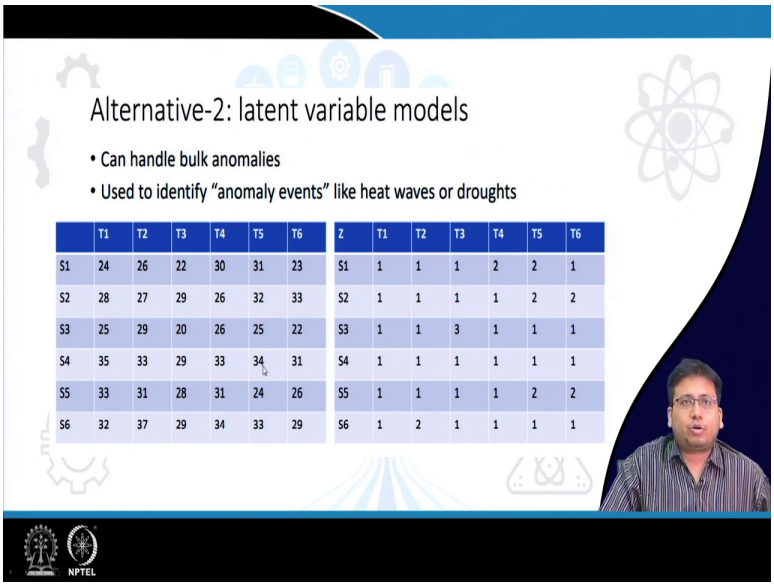


So, like this is an example of that. So, it is possible that like let us consider the 2s, the binary situation like that is anomaly or no anomaly. So, we can say that if there is; if there is no anomaly then the this distribution, I mean this variable at (s,t) it follows a gaussian distribution with

certain parameters. Let us say 30 and 10, but if there is an anomaly then it follows a different Gaussian distribution with different parameters.

Now, we have a particular observation, let us say 34.3 at X. So, now, we have to understand whether this 34.3, I mean out this value fits which of our of these models better, the is it the 1st Gaussian or the 2nd Gaussian. So, that can like accordingly the value of Z can be estimated using the Bayesian theorem. So, we know that like, so here we have expressed X as a, $(X|Z)$, the task here is to calculate $(Z|X)$ which like for which we need the Bayes theorem ok.

(Refer Slide Time: 29:17)



	T1	T2	T3	T4	T5	T6
S1	24	26	22	30	31	23
S2	28	27	29	26	32	33
S3	25	29	20	26	25	22
S4	35	33	29	33	34	31
S5	33	31	28	31	24	26
S6	32	37	29	34	33	29

Z	T1	T2	T3	T4	T5	T6
S1	1	1	1	2	2	1
S2	1	1	1	1	2	2
S3	1	1	3	1	1	1
S4	1	1	1	1	1	1
S5	1	1	1	1	2	2
S6	1	2	1	1	1	1

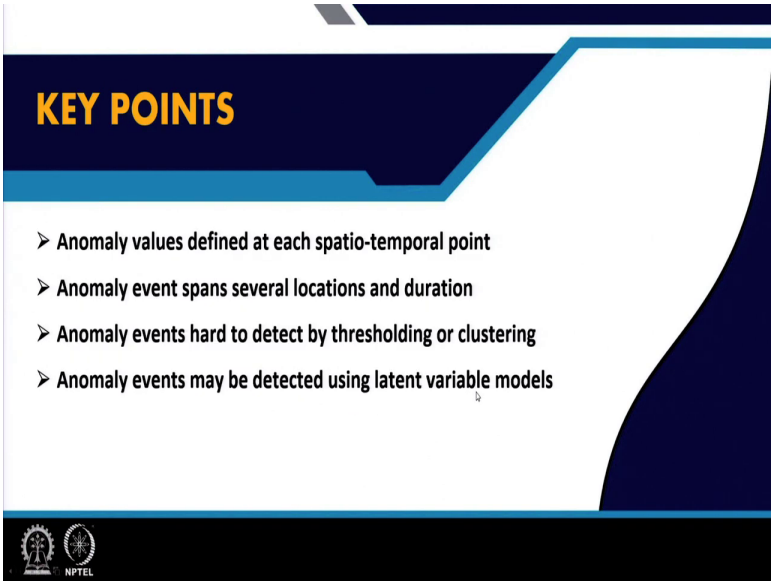
And so it is possible that once we do this kind of a analysis on the spatio temporal data set that we have, this is what we may get. So, here we are I like 1 means no anomaly, 2 means positive anomaly and 3 means negative anomaly. So, like after calculating the inference problem on these Z values, this is the; this is the map of the Z values that we get. And here we can clearly see that there is an extended anomaly event like this which starts at $S1$ at time $T4$ continues at $T5$, but it also shifts to $S2$ and it at $S2$ it continues till $T6$.

While the negative there is a negative anomaly here, which is also identified. Similarly other positive anomalies here are extended anomaly here and are isolated anomaly here, they are also identified ok.

(Refer Slide Time: 30:15)

So, that like, so this kind of this approach to anomaly detection using latent variable has been used in various research papers in the domain of climate, ones for drought detection another for identifying rainfall anomalies.

(Refer Slide Time: 30:26)



So, that message to be taken away is that the anomaly values are defined at each spatio temporal points, anomaly event spans several locations and durations. They are hard to find by thresholding and clustering, but they may be detected using latent variable models.

So, that brings us to the end of this lecture. So, we will continue our discussions on like anomalies and unusual events like this in the following lectures, till then bye.