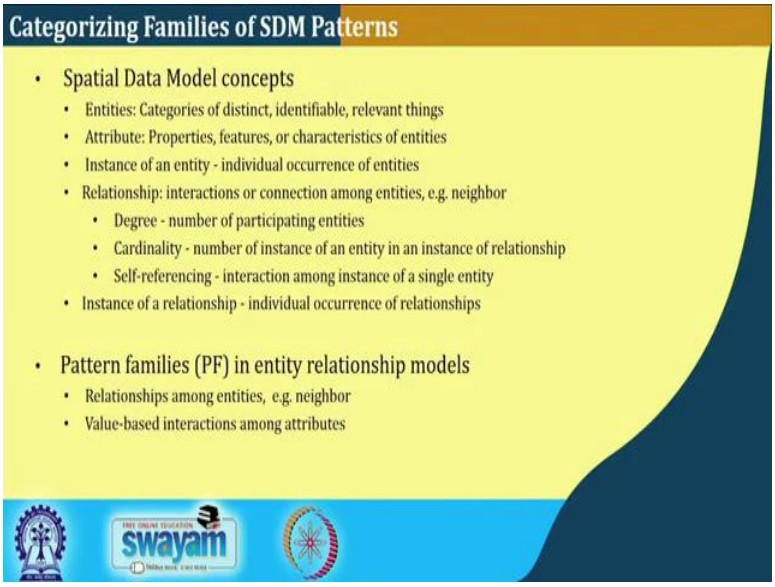
Spatial Informatics
Prof. Soumya K. Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 29
Spatial Analysis – IV

Hello. So, we will continue our discussion on Spatial Analysis, we are discussing on some aspects of spatial data mining and different aspects of spatial data mining. What are its important and how it help in doing prediction and analysis and type of things right. So, today we will continue with that discuss and see some more features, specifically that which are more predominant for spatial context rather than our non-spatial context right.

(Refer Slide Time: 00:55)



So, one is as we have as you if you recollect we discussed about patterns, spatial patterns right. So, that finding some interesting not that non-trivial patterns from the data sets right, space time data sets or spatial data sets right. So, something which is non-trivial, if it is a trivial that those patterns are not interesting right. So, those patterns are like common patterns.

So, for any if we little go back to our initial part of these codes. So, we discussed about spatial data models right. So, looking at the spatial data models again every; so, whatever we are going to do analysis, we definitely require say data model which encompasses all things.

We require definitely the spatial squares and spatial database which help us in extracting data, not only that maintaining of the whole spatial database and so and so forth. So, the spatial data model concepts is there. So, what if you just recollect there are entities, categories of distinct, identifiable, relevant objects or things.

There are attribute properties, features or characteristics of the entities like if attribute of the entities. Instances of the entities, individual occurrence of the entities and relationship, interaction or connection among the entities that is the like for example, neighbor state a or country a is a neighbor of b. So, that is a relationship right.

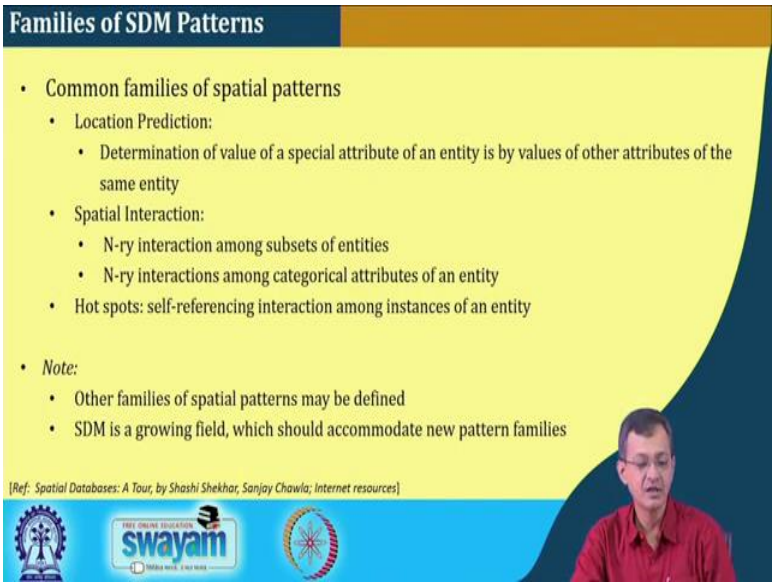
So, we have also seen different types of neighbors, we are not going into much deeper into that those already we discussed. So, there are issues of degree of when you talk about relationship, degree that number of participating entities or cardinality number of instances of a entity in an instance of relationship. Self referencing interaction among instances of a single entity right. So, instances of a relationship, individual occurrence of the relationship right.

So, instances of a relationship individual occurrence. So; that means, we have a difference instantiation of that say particular entity, particular relationship and this. So, when we talk about pattern families in entity relationship model, relationship among entities that is for example, neighbors value based interaction among the entities like how much they are related. So, I can I we want to have some quantifying.

So, I am I say that this two are more close neighbor than that to between a and b are more closer than a and c. So, that should be a way to quantify those things right. I one maybe the region a region b if I say they are more closer than this, my definition of closer maybe the distance of the centroid of the things that maybe a way of the things so; that means, the closer type of things right.

So, what is there that is data models and there are entity relationship models like to find for these pattern families in the entity relationship model. So, these are the core of the things which are the should be there.

(Refer Slide Time: 04:27)



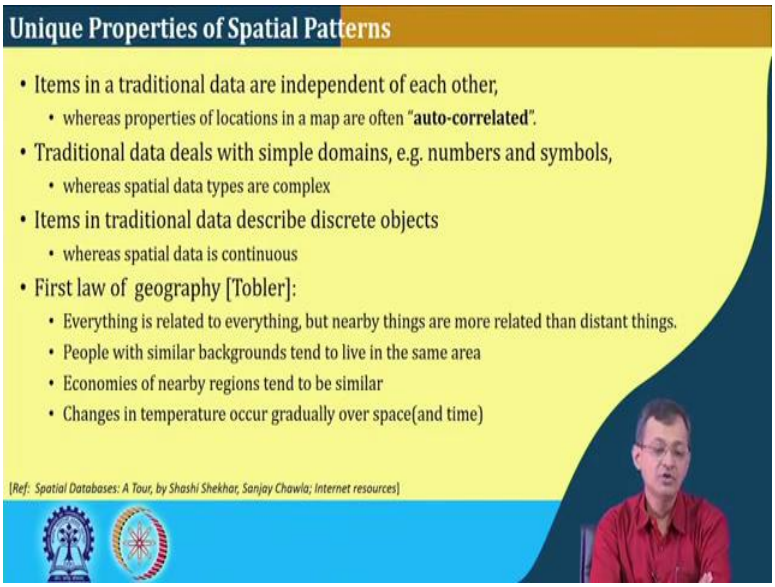
So, families also SDM patterns like Spatial Data Mining, common families of the patterns as we have seen that location prediction is one of the very predominant, spatial interaction and hotspot right. So, location prediction is determination of a value of a spatial attribute in a entity by values of the other attributes of the same entity right. I can find out that the particular event or particular activity, where it will be likely to be the next located right.

So, that is spatial attributes of an entity by a value of other attributes of the same entity I want to predict that location. Spatial interaction is a N-ry interaction among the subset of the entities like the if there is a interaction. So, it can be N-ry interaction, N-ry interaction among categorical attributes of an entity right. I can instead of subset of the entities, there can be a categorical interaction among the category and hotspot is self referencing interaction among the instances of an entities right.

So, you want to find out that particular hotspot like we are looking for crime hotspot etcetera. So, the number of occurrences are much more in that particular crime and region to the in a particular region then the in other regions. So, something may be noted that other families or patterns may be defined like it based on your requirement you can define other families of pattern or particular domain requirement or application requirement.

And, SDM per se is a growing field we should accommodate new families or patterns. So, that so to say that the spatial data mining framework should be able to accommodate this new families of pattern.

(Refer Slide Time: 06:37)



So, unique properties of spatial patterns. So, there are some of the unique properties of spatial patterns which basically distinguish it from other non-spatial patterns. Like items in a traditional data are independent of each other right, items in a traditional data sets are independent of each other right. Like say if you have a student database right; so, or so say even different tipples I cannot see there is a inherent relationship. They may be in the same class and type of things, but the different values of the item sets are independent of one another right.

Like marks of a student is independent of the marks of the other students right, that we may try to find out with respect to that there is a teaching qualities high and type of things and like that or bank account numbers right. The amount of balance of a one customer is independent of the balance of the other customers right. So, there is what we say items in a traditional data sets are independent of each other right whereas, properties of a location in a map are often correlate, auto correlated right. So, there is a inherent thing called auto correlated.

The same thing correlated to the same type of object right like temperature of a region cannot be independent of the temperature of the nearby region alright. So, it is there is

influence on the thing or rainfall of a region is somewhat dependent on the rainfall of the other region right of the nearby region right. Temperature of Kharagpur may not be directly influenced by the temperature of say somewhere in Pune or somewhere in Bangalore. But, temperature of Kharagpur is definitely have some relation of that of the Midnapore.

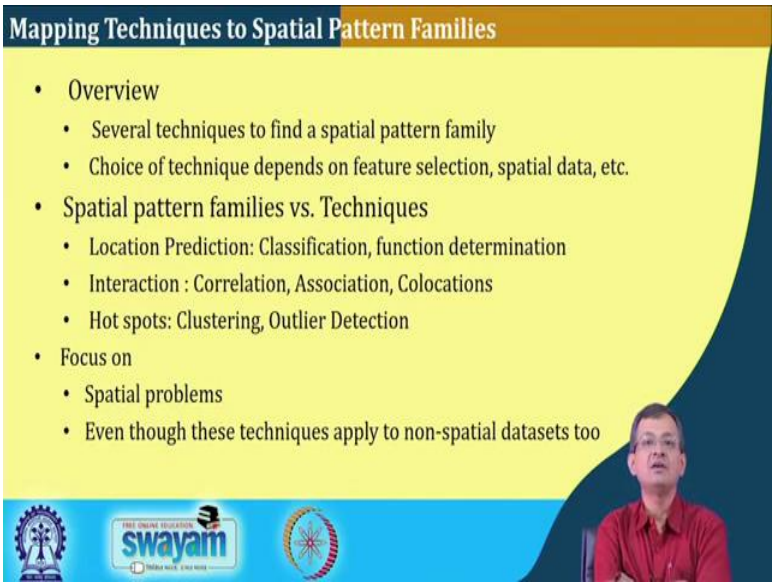
So, this temperature thing is auto correlated right. So, traditional data sets deals with simple domain like numbers and symbols right whereas, spatial data types are more complex right, it has a shape and other type of thing. So, there is no not simple domains. Items in a traditional data described discrete objects right whereas, spatial of data sets are mostly continuous right. So, there is a there are differences and there is a first law of geography given by Tobler. He says everything is related to everything, but nearby things are more related than distance things right.

So, any phenomena which is any spatial phenomena or any geography or geospatial phenomena are related to one another right, but the nearby things are more related than the distance thing right. So, that is why we say that the temperature of Kharagpur is influenced or is more related to the temperature of the nearby places right. So, it is or different type of thing, something some sort a if some seasonal disease is happening in Kharagpur; it is likely that nearby things, it is not that the far things. But, that maybe some relation, but there is more you go the distance the relationship is become weaker.

So, people with similar background tends to live in the same area right like we heard about hear about that HIG flat, MIG flat, LIG flat low income group, High middle income group etcetera. So, why they are met in a particular blog or type of things because, it is seen that peoples with the same background or same income group are trained to in the same area.

Economy of nearby region tends to be similar right. So, that it is there, changes in temperature occur gradually over space time right. So, if we say the temperature of a particular region, it is gradually over the space it changes right and also over gradually over the temporal scale also changes. So, these are things which comes out of that first law of geography by Tobler.

(Refer Slide Time: 11:11)

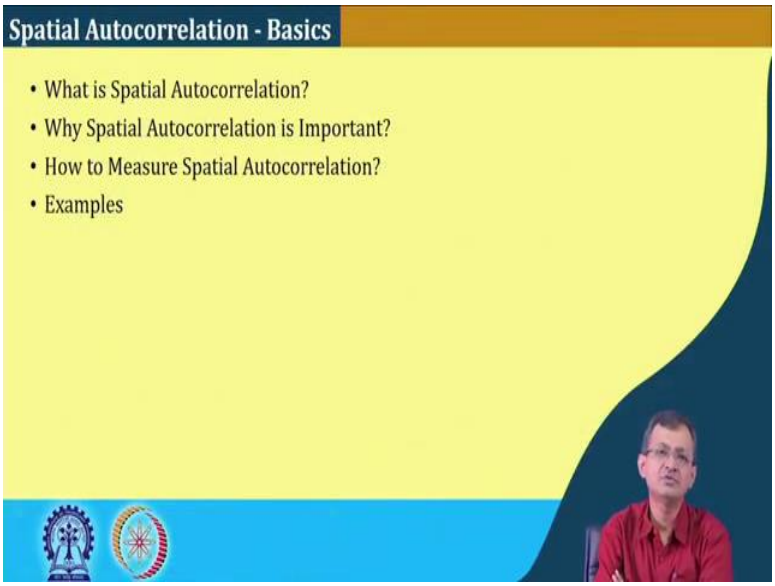


So, we have mapping techniques to spatial pattern families. So, that several techniques to find out the spatial pattern families right, how to identify these spatial patterns. So, it is not that straightforward or not that we can use that state way all the techniques which are there in the traditional pattern analysis or non-spatial data analysis right. Choice of technique depends on the feature selection spatial data etcetera so, which techniques independent.

Spatial patterns families versus techniques, location prediction like for location prediction we use typically classification, function determination and type of things. Interaction what we will do correlation, association, colocation type of things right. When we look about interactions of the things, when we find out the location then we want to classify and function determination that what should be the next location and type of things. When we look about the hotspot, we look for clustering outlier technique that the things like that, when we look at the hotspot type of things right.

So, but we see that finding different patterns techniques we need to deploy different type of techniques or methods or approaches right. So, here the focus is on our this spatial problem right, even though this techniques apply to non-spatial data set also. So, whatever the techniques we are which are talking about these are applicable; many of them are applicable to non-spatial data set also, but here we mostly try to deal with spatial data into consideration.

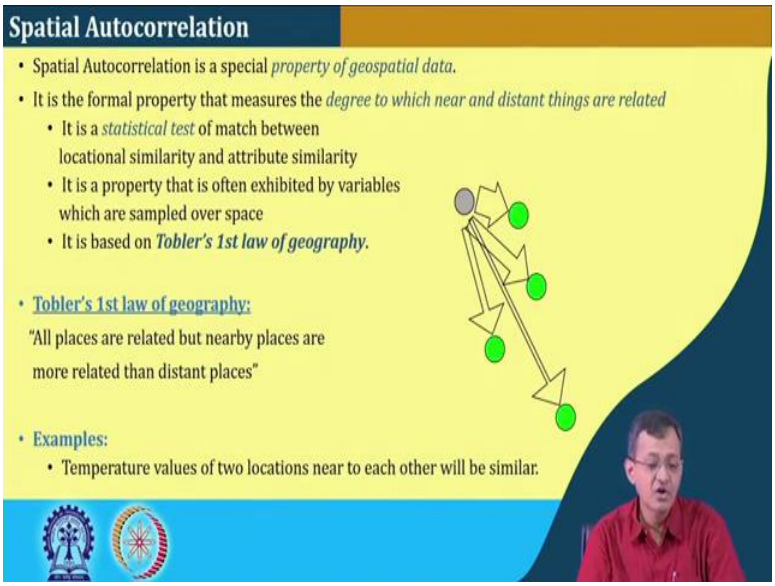
(Refer Slide Time: 13:05)



So, one things which is coming now and then what we say spatial auto correlation right. So, before we go other type of techniques, we thought that it will be nice that we can have a overview of the spatial autocorrelation right. So, what does it mean? What are the typical or standard techniques to calculate? There are different type of techniques and there are different mechanisms to do that, we want to see that what are the different, how good can be calculated.

So, we want to see that what is the spatial autocorrelation per se, why spatial autocorrelation is important for our all these spatial analysis and stuff like that. How to measure spatial autocorrelation maybe one or two examples.

(Refer Slide Time: 13:55)



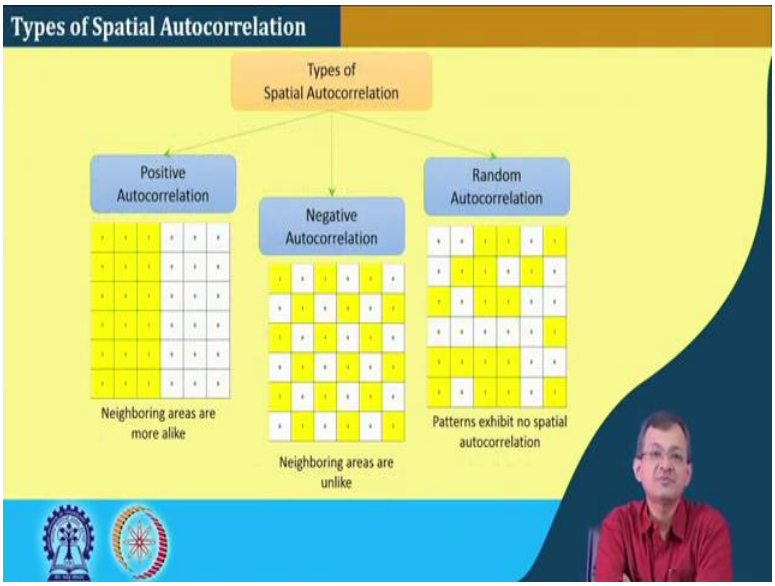
So, what when we talk about spatial autocorrelation is a spatial property of geospatial or spatial data right. So, spatial autocorrelation is a typical property of spatial, how to say geospatial data. It is a statistical test of a, it is a formal property that measure the degree to which the near and distance objects are related. So, it is a formal property that measure the degree to which the near and distance things are related right. So, it is a statistical test of match between location similarity and attributes similarity right so, this is important.

So, it is a statistical test between the location; that means, the where the coordinate wise things are there and the attribute value similarity right. So, it is a thing and it is a property that is often exhibited by variables which are sampled over space. So, and it is a property what we were we sample the things over space, it comes into play. As in our spatial context this is very much there, the things has sampled over space. So, this autocorrelation come into play and it is based on as we have seen in the last slide Tobler's 1st law of geography right.

So, just to repeat the basic philosophy of the Tobler's 1st law of geography tells that all places are related, but nearby places are more related to then distance places. So, what it says that the all places are related, but the nearby places are more related than the distance places right. So, there is if we look at temperature those will be more related which are next to make next to the region now, but definitely there may be relations with the distance things also.

Example like temperature values of two location near to each other will be similar, will be more similar or more related that the temperature of two locations at a faraway distances. So, this is the my if we in the things which I am comparing these are the different things. So, these are more closely and these are far away.

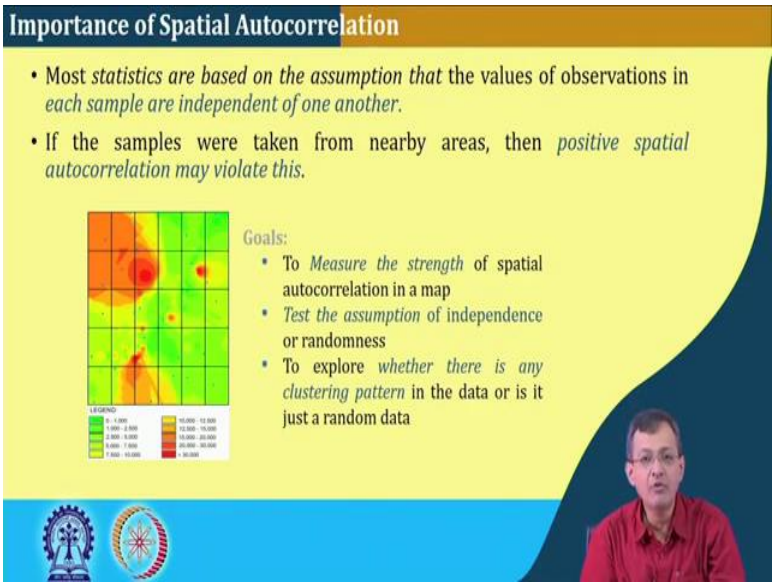
(Refer Slide Time: 16:15)



So, spatial if you look at the spatial type of auto spatial autocorrelation like types of spatial autocorrelation. So, one is positively correlated right so; that means, it is in a neighboring areas are more alike and type of things or it can be negative regurgitated, like neighboring areas are more unlike here if you see rather than and it can be a random correlation right, pattern exhibit no spatial auto correlation right.

So, there is no spatial auto correlation. So, what these are things are not you cannot predict that what is going on which are it is nearby etcetera. So, this is a positive correlation, negative correlation and random autocorrelation. So, these are the three type of spatial autocorrelation traditionally.

(Refer Slide Time: 17:21)

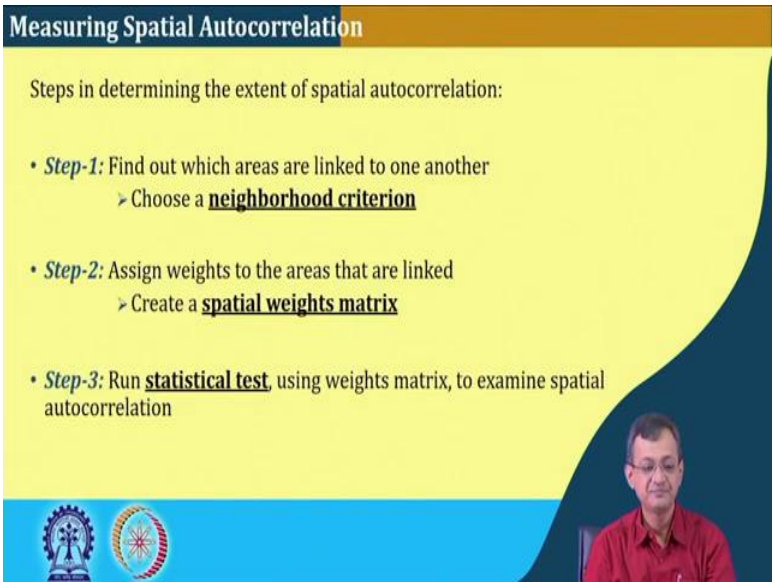


Now, why this spatial correlation is important? Most statistics are based on assumption that a values of observation in each sample are independent to each other right. Most statistics are based on the assumption that the values of observation in each samples are independent to each other right. If the samples were taken from nearby areas then positive spatial autocorrelation may violate this right.

So, though the statistics based on the assumption that the values of the observation each sample are independent. When we take samples in typical statistical analysis, these observations are independent. Whereas, if the samples were taken from nearby areas then positive spatial autocorrelation may violate this. So, like this is a particular map of the things and then we say that to measure the strength of spatial autocorrelation in a map, we want to given a map we want to find out that what is the how much auto related things are there.

Test the assumption of independent or randomness to explore whether there is any clustering pattern in the data or it is just a random data right. So, this is our basic goal seen and if you if we recollect that our few slides back or the previous lectures these other things which dictates the how this SDM patterns will generate right. So, this is important. So, autocorrelation makes as a foundation for doing any type of spatial analysis or spatial data mining type of activities.

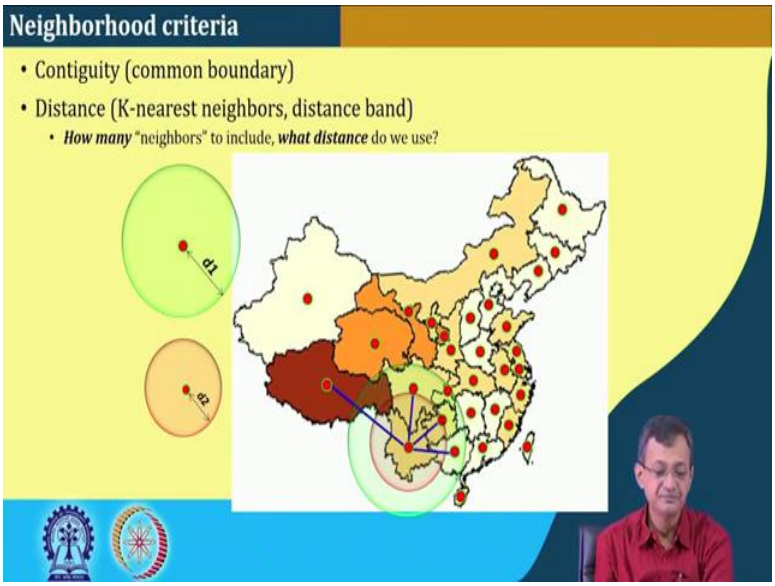
(Refer Slide Time: 19:05)



Measuring spatial autocorrelation, there are typically 3 steps. Find out which areas are likely are linked to one another, choose a neighbourhood criteria. So, which areas are linked. Secondly, assign weights to the areas that are linked create a spatial weight matrix and run statistical test we using weight matrix to examine spatial autocorrelation right.

So, one is that doing a choosing a neighborhood region, neighborhood criteria, assign weights to the areas that are linked that create a spatial weight matrix and run statistical state test using weight matrix to examine spatial autocorrelation. So, these are the things which are there.

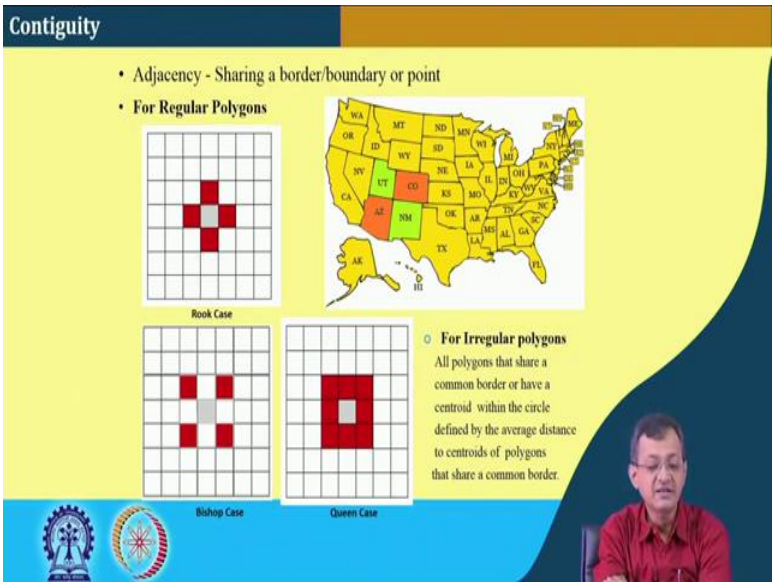
(Refer Slide Time: 19:59)



So, neighborhood criteria; so, it is contiguity common boundary maybe a neighborhood criteria. Distance K nearest neighbor distance band how many neighbors to include, what distance do we use and type of things right. So, like here if you see if particular this region; so, these are my distance from the say centroid of the different neighborhood region.

Now, I need to define that if I my d 2 is my consideration is my neighbor then maybe one or two maybe touching these are the things, where if it is d 1 is my neighbor then a few more neighbor right. So, it looks at that defining the neighborhood criteria is important that how we want to look at the things.

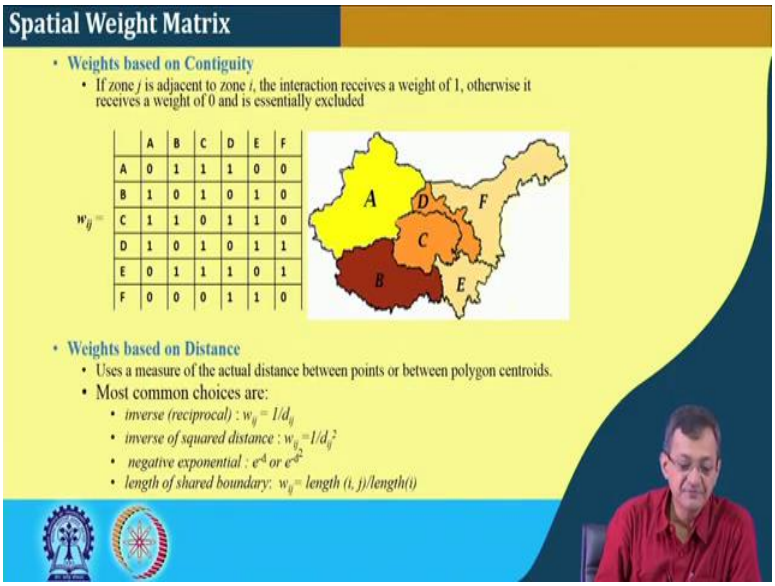
(Refer Slide Time: 20:53)



Contiguity adjacency sharing a border or boundary or point right. So, adjacency is things for a regular polygon, it can be this sort of things or those who are accustomed or those who have played chess they know that these are the things the movement of a particular rook right, either vertically or horizontally. So, this type of things if it is my search space is just these four then we say this is a rook case right. So, I want to find a neighbor considering this rook case or type of things and in a again chess board the bishop movements is diagonal.

So, I can have a bishop case or it can be a queen case right, where given a point all surroundings are the queen can moved in all things either diagonal or vertical or horizontal. So, this is queen case. So, for irregular polygon all polygons that share a common boundary border or have a centroid within that particular circle of search or circle defined by the average distance up to the centroid of the polygon that share a common boundary right. So, that can be the things right. So, we can have different type of scenarios like this.

(Refer Slide Time: 22:17)



Now, spatial weight matrix, weight based on contiguity if a zone j is adjacent to a zone i interaction receives a weight 1 otherwise it receives a weight 0 and essentially excluded. Like A to A there is no as such we are not considering that they are thing the same region. So, 0 A to B connection is there, A to C is there, A to D is there, A to E no, A to F no. So, there is no contiguity and similarly I can create like C to A is there, C to B is there, C to C is not considered, C to D, C to 1 and so and so forth right.

So, this is the different what we say weight based contiguity or given 1 0 type of values in the line. So, weight based distance you we can have a different type of matrices like use a measure to actual distance between the points or between the polygon centroids right. If it is a points or polygons centroids find the actual distance. So, there can be different way of calculation: one is that inverse proportional w ij equal to 1 by d y that will be captured like d ij; that means, if I say these two centroid this and this. So, their weight is calculated the inverse distance right.

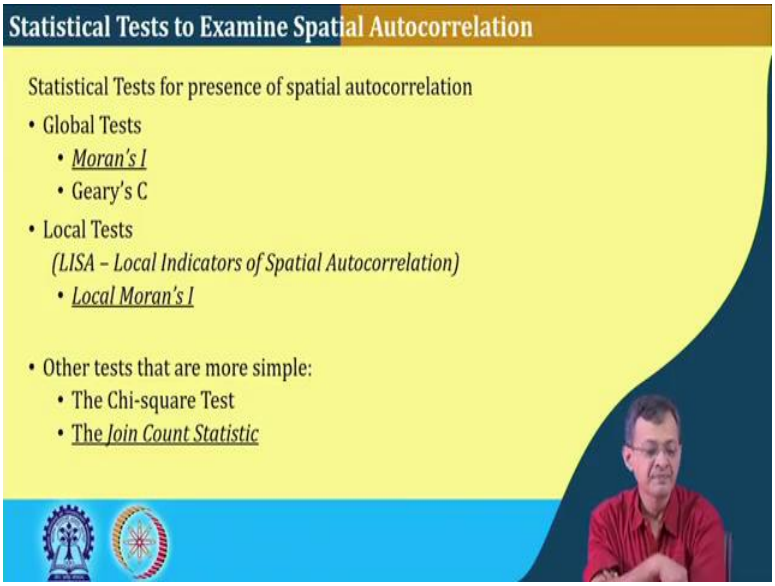
So, in other sense if it is if we consider this weight for influence so, more the distance it is the influence will be less right. So, inverse distance is a very popular, inverse of square distance is also popular instead of only d ij 1 by d ij square. Negative exponential it is 1 minus d, it is 1 minus d square or length of sharing boundaries right. So, I can have a length of the sharing boundary. So, length ij divided by length i can be the thing right; so, this can be different way of matrices. So, what we are trying to see here? There

spatial weight matrix and try to see that how I can have a quantitative value of this matrices which can be used for my spatial analysis.

Now, I what I am to find out that how these things influences, if there is a effect here how it will be affected there. If there is a congestion in a region A of a say part A or the a particular region of a city, how that will be affected in the other region of the cities right. So, how it influences the things right. So, one way is that if I see this type of data and distances etcetera, I can try to predict type of things right.

So, even there are different type of parameters may come into play, that how the regions are connected if it is a traffic movement. If it is a say flooding of the things right the how things will be there right. There are different patterns, different parameters like elevation and rainfall level and type of so and so forth. Nevertheless, things are I want to calculate these some weight based on the distances.

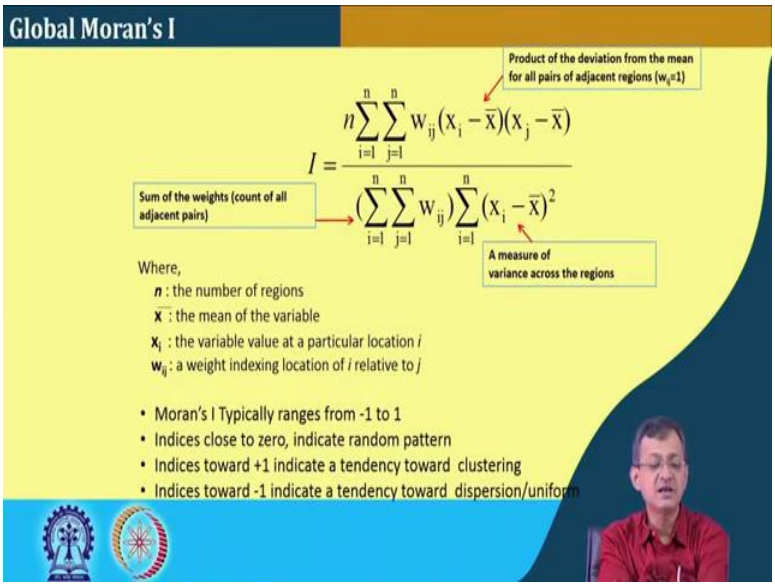
(Refer Slide Time: 25:51)



Now, there are several statistical tests to examine spatial autocorrelation right. So, we have various statistical test for autocorrelation. The statistical test for performance of spatial autocorrelation, there are some of the global test like taking the whole of the region. One is Moran's I one of the very popular and then other is here you see this is another scheme and local test there is a test called LISA: Local Indicators of Spatial Autocorrelation.

So, local Moran's I is the local test. There are other test like chi square test, join count statistics and type of things. Out of these what we see that Moran's I, local Moran I, Moran's I and join count statistics, these are popular and can be applied to spatial datasets we are dealing with right. So, this is the things we are looking at.

(Refer Slide Time: 26:55)

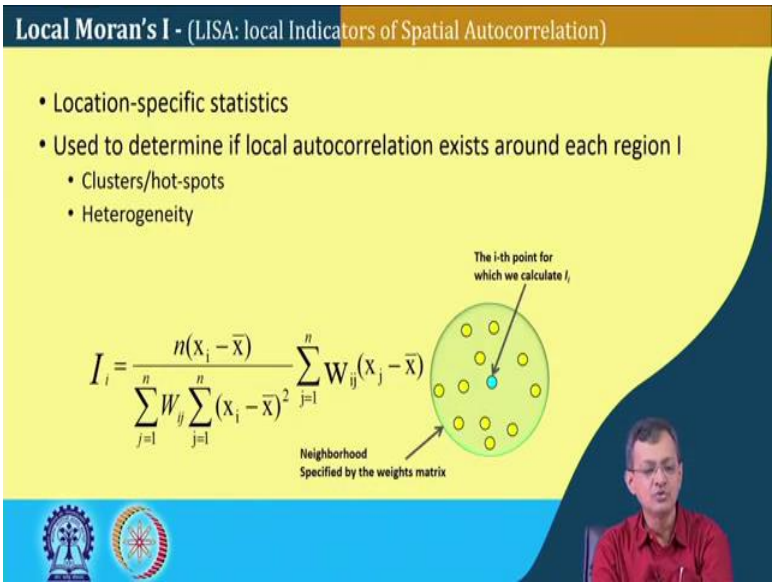


So, this is global Moran's I expression, how to calculate that is the product of the deviation from the mean for all pair of adjacent regions right. So, a measure of variants across the region. So, this is thing or sum of weights counts of all the adjacent regions that can be these are the different components. So, the Moran I is calculated in this fashion. So, n equal to number of regions, X bar the mean, X j is the variable value in the particular location, j Xi and W ij is the weight index location from of i relative to the j right. So, these are the different parameters.

So, Moran I typically ranges Moran's I typically ranges from minus 1 to plus 1, indices close to zero indicates random pattern. So, if it is a close to zero, then it indicates random pattern; indices towards plus 1 indicate a tendency towards the clustering right. So, if it is a if the Moran's I value is moved towards plus 1 so, there is a tendency of clustering is there. If it is towards more towards minus 1 indicate tendency towards dispersion or some sort of a uniform distribution type of things right. So, there is no specific inherent clustering into the things.

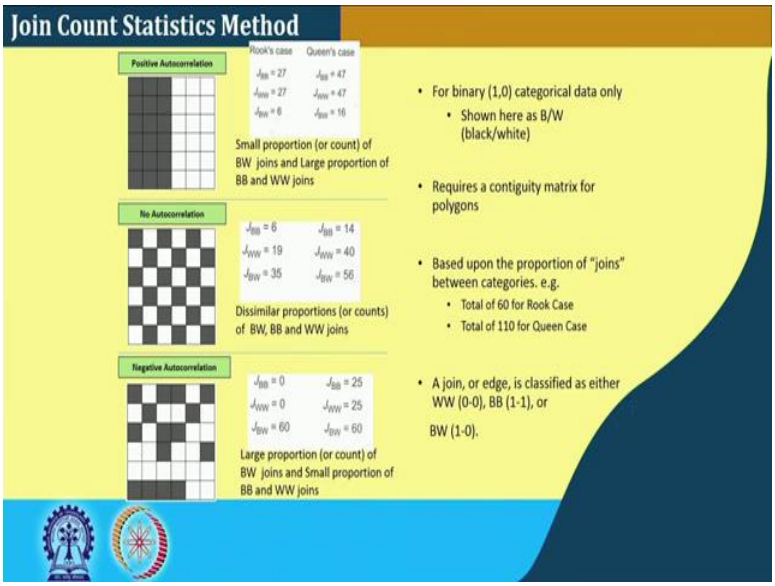
So, if the value is towards plus 1 then there is a inherent clustering right. So, indices towards minus 1 indicates that is a dispersion or uniform uniformity into the things, that may not be very interesting depends on that how what the application area is there. So, this is a one measure to look at the things.

(Refer Slide Time: 28:55)



So, there is a local Moran I Moran's I which is some variant of the things, local specific statistics use to determine if the local autocorrelation exist around each region I. So, cluster clusters hotspots or heterogeneity so, it is there. So, it defines in nearby region right. So, its defines which a particular buffer zone. So, I is the point where we calculate I, neighborhood specified by the weight matrix; so, it is the neighborhood zone. So, it is this is where in this case is a global taking all consideration, it is a local taking the consideration of a particular neighborhood region.

(Refer Slide Time: 29:37)



And we have a joint count statistics for things. So, there can be different type of variation, one this is a for positive autocorrelation and the if there is a fully uniform or spread like this, there is no autocorrelation and if it is there is a negative autocorrelation right. So, there are three type of things, here it is a sort of a binary image, we can calculate different rook case; so, this is with the rook case and queen case. So, different type of values calculation we can do.

Like here what we say joint of BB, there is black to black is 27, white to white is 27 and join count statistics of black and white is 6 right, that is easy to calculate. So, 27 is if you see that how much it is black to black sharing is there say 27; 1 2 3 4 5 6 7 8 9 10 11 12 and 1 2 3 4 5 5 into 3 15, 15 plus 12 is 27 right where, 6 6 you see 1 2 3 4 5 6 is the things. Similarly, if this is the rook case, can rook can go only in the horizontal and vertical direction.

Similarly, we can calculate it by Queen's case and for all these autocorrelations right. So, for binary 1 0 category data only shown in black and white requires contiguity matrix for polygon, based upon puts in of joins between the categories for the total 60 for rook case, 110 for Queen's case and type of things. A join age is classified as either WW 0 0 or black black 1 1 or black W 1 0 type of things right. So, these are the way we calculate.

So, for today's discussing what we have seen that we tried to look at that how the spatial patterns can be quantified, more specifically we try to see that one of the important

property of spatial data mining or any spatial analysis is the spatial autocorrelation. So, how what it is, why it is important, how to calculate; we will be continuing this discussion in our subsequent lecture.

Thank you.