

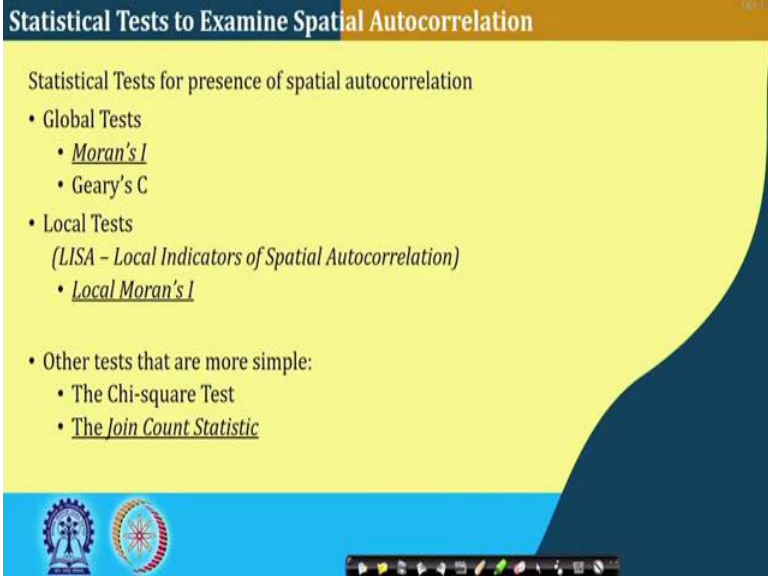
Spatial Informatics
Prof. Soumya K. Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 30
Spatial Analysis – V

Hello. So, we will continue our discussion on Spatial Analysis, on the part in our course on Spatial Informatics, as we have discussing for last cup few lectures. So, we have seen that there are we are dealing with different aspects of the things starting from we seen that spatial data, warehousing, data mining, the normal of means standard warehousing, traditional data, warehousing, mining and then the spatial data mining aspects then looked into different other aspects of spatial analysis, right, some sort of things which falls into spatial computing aspects also.

So, rather in the last lecture if you recollect we are discussing about spatial autocorrelation. So, this is a spatial property of spatial features where there is a correlation between the same entities, right. Like, things like the temperature of a region is highly correlated of with the temperature of the nearby regions. So, this is a feature which is this sort of features which are manifested in the spatial data. So, it is important to work on to see that what is the spatial autocorrelation, rather autocorrelation plays the important role in different aspects or different analysis and computing aspects of a spatial data.

(Refer Slide Time: 01:38)



The slide is titled "Statistical Tests to Examine Spatial Autocorrelation" in a blue header bar. Below the title, the text "Statistical Tests for presence of spatial autocorrelation" is displayed. The slide lists three categories of tests: Global Tests (Moran's I, Geary's C), Local Tests (LISA - Local Indicators of Spatial Autocorrelation, Local Moran's I), and Other tests that are more simple (The Chi-square Test, The Join Count Statistic). The slide has a yellow background with a blue wavy shape on the right side. At the bottom, there are two circular logos and a small navigation bar.

Statistical Tests to Examine Spatial Autocorrelation

Statistical Tests for presence of spatial autocorrelation

- Global Tests
 - Moran's I
 - Geary's C
- Local Tests
 - (LISA - Local Indicators of Spatial Autocorrelation)
 - Local Moran's I
- Other tests that are more simple:
 - The Chi-square Test
 - The Join Count Statistic

So, just to continue our spatial autocorrelation. So, there are some statistical test to examine spatial autocorrelation. Again, so to say these are not these tests are not from suddenly discovered perform, these are already there many of you might have already read this thing. So, we are what we are doing? We are using those things in our spatial context. Like, one is a global test where Moran's I and this Geary C are the two major thing.

Local test which is a local Moran I, Moran's I and there are other tastes like chi square test and joint statistics. Out of that Moran's I, local Moran's I and joint count statistics are popular or we will see a quick overview of those things. Others there are, other different type of measures which you can follow any spatial analysis book or even processes is a course book to refer to that what are the different things, right.

(Refer Slide Time: 02:38)

Global Moran's I

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\left(\sum_{i=1}^n \sum_{j=1}^n w_{ij} \right) \sum_{i=1}^n (x_i - \bar{x})^2}$$

where,

- n : the number of regions
- \bar{x} : the mean of the variable
- x_i : the variable value at a particular location i
- w_{ij} : a weight indexing location of i relative to j

• Moran's I Typically ranges from -1 to 1
 • Indices close to zero, indicate random pattern
 • Indices toward +1 indicate a tendency toward clustering
 • Indices toward -1 indicate a tendency toward dispersion/uniform

So, global Moran's I it is more of a statistical measure rather local is also statistical measures where here we see that the we have this product deviation from the mean from all like x_i , \bar{x} , x_j from the \bar{x} that what is the mean deviation and this is variance calculation, right, weighted variance calculation sum of weight count all adjacent pairs and weighted calculation. And so, a number of regions mean of a variable, n is the number of region, mean of a variable, and various variable this x_i is the so to say that the random variable, value in a particular location i and w_{ij} weight indexing location i relative to j . So, these are the things, right.

So, what we are trying to see that whether these adjacent areas that how far it is there and what are the influences and try to calculate the things. So, Moran's I value typically ranges from minus 1 to 1. Indices close to zero, indicates random pattern. Indices close to plus 1 indicates towards clustering, they group together. Indices towards minus 1, when the value is towards minus 1 indicate tendency towards dispersion or uniform. So, that is a either there is a dispersed thing or it is uniformly distributed, right. So, this gives a way that how things are together.

(Refer Slide Time: 04:13)

Local Moran's I - (LISA: local Indicators of Spatial Autocorrelation)

- Location-specific statistics
- Used to determine if local autocorrelation exists around each region i
 - Clusters/hot-spots
 - Heterogeneity

$$I_i = \frac{n(x_i - \bar{x}) \sum_{j=1}^n W_{ij}(x_j - \bar{x})}{\sum_{j=1}^n W_{ij} \sum_{j=1}^n (x_j - \bar{x})^2}$$

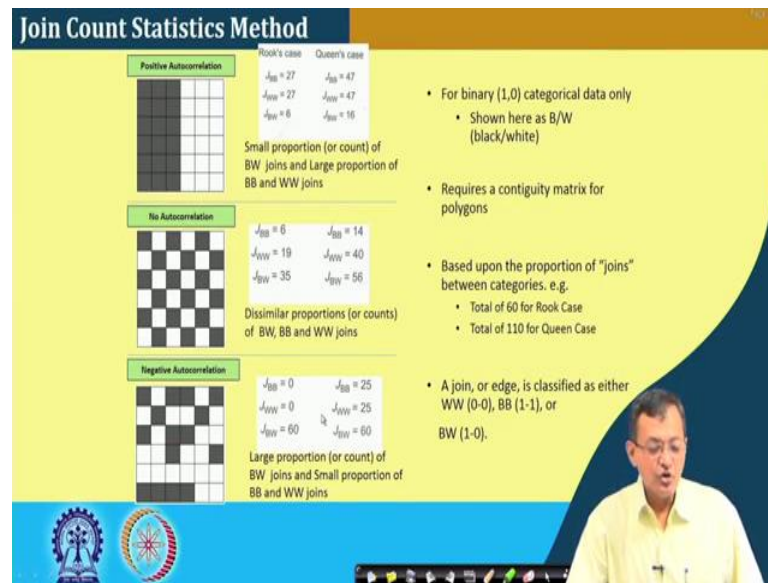
The i -th point for which we calculate I_i

Neighborhood Specified by the weights matrix

And local Moran's I is it is more of a location specific, right local Moran's I is a local indicator for spatial autocorrelation. So, it is used to determine if the or local autocorrelation exists around a region around each region i . So, within a particular specified neighbourhood region, right.

So, this is more amicable for finding clusters and hotspots, heterogeneity within that particular buffer zone, all right. So, the we calculate in the same way except that here all summing up all the things we take only this in number of things within that particular region, right and that rate matrix. So, this is more of a with respect to a X_i of the things with this particular location and calculate the thing with respect to it.

(Refer Slide Time: 05:20)



So, we have a joint statistics method there is another popular things as if you recollect we discussed about rook case, queen case and bishop case, right. So, 3 type of things. So, in case of a rook case it is horizontal vertical line, in queens case it is all surrounding and in case of Bishop this is diagonal lines, right.

So, this is finding out that what are the how it is rated. Like here in this case if you say this join of B to B black to black there are 27 such surfaces or a sorry such boundaries, right. Like here 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and these 3 into 1, 2, 3, 4, 5, 6, so 6. What we are having here? 1, 2, 3, 4, 5, 5 boundaries, 5 into 3, 15 plus 12, 27, right. Similarly, I can calculate all the things.

So, it gives a small portion or count of black and white joints, large portion of black and white joints in a binary things, right. Whereas, this gives a positive correlation, if this sort of things are there. If it is no correlation when this sort of situation are there and if for negative correlation we have this sort of situations, right. This similar portion of or counts for black BW, BB and WW, right.

And where here large portion or count on the black and white or two different class and smaller portion on BB and WW, based on that we can have a by this count we can have a joint count statistics that where is positively correlated, negatively correlated or no correlations are there, right. So, correlation becomes a important factor to finding out that what are the different relationship between the feature sets.

(Refer Slide Time: 07:34)

Join Count Statistic: Calculation

Test Statistic is given by: $Z = \frac{\text{Observed} - \text{Expected}}{\text{SD of Expected}}$

Expected value is given by:

$$E(J_{BB}) = kp_B^2$$

$$E(J_{WW}) = kp_W^2$$

$$E(J_{BW}) = 2kp_Bp_W$$

Standard Deviation (SD) of Expected is given by:

$$SD(J_{BB}) = \sqrt{kp_B^2 + 2mp_B^3 - (k + 2m)p_B^4}$$

$$SD(J_{WW}) = \sqrt{kp_W^2 + 2mp_W^3 - (k + 2m)p_W^4}$$

$$SD(J_{BW}) = \sqrt{2(k + m)p_Bp_W - 4(k + 2m)p_B^2p_W^2}$$

Where: k is the total number of joins (neighbors)
 p_B is the expected proportion Black
 p_W is the expected proportion White
 m is calculated from k according to:

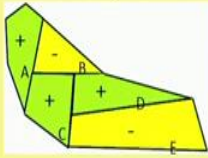
$$m = \frac{1}{2} \sum_{i=1}^n k_i(k_i - 1)$$

Joint count statistics calculation. Test statistics given by that Z value observed value by expected value divided by the standard deviation of the expected value, all right. And we have this standard formulation to calculate such statistics, right. So, the total number of joint expected portion, expected portion of the white it is calculated from the other things. So, we calculate this joint statistics binds.

(Refer Slide Time: 08:05)

Example: Moran's I

Research question: Is the areal pattern of + and - values randomly distributed amongst the polygons?



Polygon	Value
A	20
B	10
C	15
D	16
E	9
Mean	14
Std. Dev.	4.53

$$W = \{w_{ij}\} = \begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} & \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \end{matrix}$$

$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{(\sum_{i=1}^n \sum_{j=1}^n w_{ij}) \sum_{i=1}^n (x_i - \bar{x})^2} = -0.2806$

For $i=C$, $I_i = \frac{n(x_i - \bar{x})}{\sum_{j=1}^n w_{ij} \sum_{j=1}^n (x_j - \bar{x})^2} = 0$

I value is less than 0. Therefore, the areal pattern may be dispersed. Z-test is required.

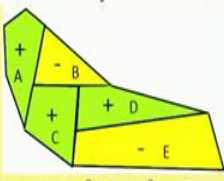
So, say in a say this typical scenario weather is a real pattern of plus and minus, values randomly distributed amongst the polygon, right. So, polygon A 20, polygon B and so on

and so forth, right. So, and we have these matrix, like A to A there is, this is the diagonal is 0, A to B is 1, A to C is 1, B to A is 1, C to B and like that we have a that connectivity matrix, right. And then we calculate these morons I calculation. And I value less than 0 therefore, the aerial pattern may be dispersed, so Z test is required to find out that whether there is a any appropriate correlation, right.

(Refer Slide Time: 09:03)

EXAMPLE: JOINT COUNT STATISTIC

Research question: Is the areal pattern of + and - values randomly distributed amongst the polygons? $Z_{crit}=1.96$




We now have **three types** of boundaries:

- ++ boundaries (2)
- +- boundaries (5)
- -- boundaries (0)

Given, $p[+] = \frac{3}{5}, p[-] = \frac{2}{5}$ Therefore, $E[+-] = 2 \times 7 \times \left(\frac{3}{5}\right) \times \left(\frac{2}{5}\right) = 3.36$

$m = 28$

$V[+-] = \left(2 \times (7 + 28) \times \left(\frac{3}{5}\right) \times \left(\frac{2}{5}\right)\right) - \left(4 \times \left(7 + (2 \times 28)\right) \times \left(\frac{3}{5}\right)^2 \times \left(\frac{2}{5}\right)^2\right) = 2.28$



So, another thing, same for the things if Z critical is 1.96 and if we calculate these Z values. And we get that if the Z test is less than the critical value then we do not reject that hypothesis the pattern is random. So, it is Z test is less that critical value, right.

(Refer Slide Time: 09:14)

Example: Joint Count Statistic(contd.)

1. $H_0: O[+-] \sim E[+-]$ (Areal pattern is random)
2. $H_A: O[+-] \neq E[+-]$ (Pattern is spatially autocorrelated)

We will calculate the test statistic using:


$$Z_{\text{test}} = \frac{(\text{Obs. "++"}) - E[+-]}{\sqrt{V[+-]}}$$

O : Observed Value
 E : Expected Value
 V : Variance

$$= \frac{5 - 3.36}{\sqrt{2.28}}$$
$$\approx 1.086$$

$Z_{\text{test}} < Z_{\text{crit}}$

Therefore, we don't reject H_0 : "the areal pattern is random"




So, this is fully statistical approach which are already there in the literature. We are using those to find out that what is the how these things are correlated, right. So, this scenario may not be like that hypothesis, the aerial pattern is random we do not reject that hypothesis, right. So, that is no apparent correlation out it there.

(Refer Slide Time: 09:54)

Location Prediction

- Classical method:
 - Logistic regression, decision trees, bayesian classifier
 - Assumes learning samples are independent of each other
 - Spatial auto-correlation violates this assumption!
 - Map display where the properties of a pixel is independent of the properties of other pixels?
- Spatial methods
 - Spatial auto-regression (SAR),
 - Markov random field
 - Bayesian classifier




So, given this autocorrelation and based on those things we have some of the other things like one is that whether we can have a location prediction. So, classical method logistic regression, decision tree, Bayesian classifier, these are the some of the classical methods

which assume that landing samples are independent to each other, right which is not true for in our case or in spatial domain, right. Spatial autocorrelation violates these assumption, right. So, these are not like temperature of Kharagpur is not independent of the temperature of IIT, Kharagpur is not independent of the temperature surrounding IIT, Kharagpur, right. So, there is a autocorrelation between this thing.

Whereas, in case of our classical methods we take this variable X to be independent X or X_i that instances are independent to each other. So, map display where the properties of a pixel is independent of the properties of other pixels, right. So, there are different spatial methods, spatial autoregressions are Markov random field, Bayesian classifier and type of things, right. They are they are different spatial method which are being employed.

(Refer Slide Time: 11:07)



Association Rule Mining

- Classical method:
 - Association rule given item-types and transactions
 - Assumes spatial data can be decomposed into transactions
 - However, such decomposition may alter spatial patterns
- Spatial methods
 - Spatial association rules
 - Spatial co-locations
- Note: Association rule or co-location rules are fast filters to reduce the number of pairs for rigorous statistical analysis, e.g correlation analysis, cross-K-function for spatial interaction etc.

So, there is another aspects we briefly talked about it, there is a association rule mining. Like how a item set or I can say their spatial event is associated rather events like. So, I say a set of events a or a item set a implies b , right. Like, we are saying that say if I say that if there is a road blockage there will be congestion which some of the things are trivial we know. But if the road blockage of region a creating a congestion in region b that may not be apparently trivial or something which is happening some climatic phenomena in some region when a faraway region something is influence that may not that may be interesting rule to this, right.

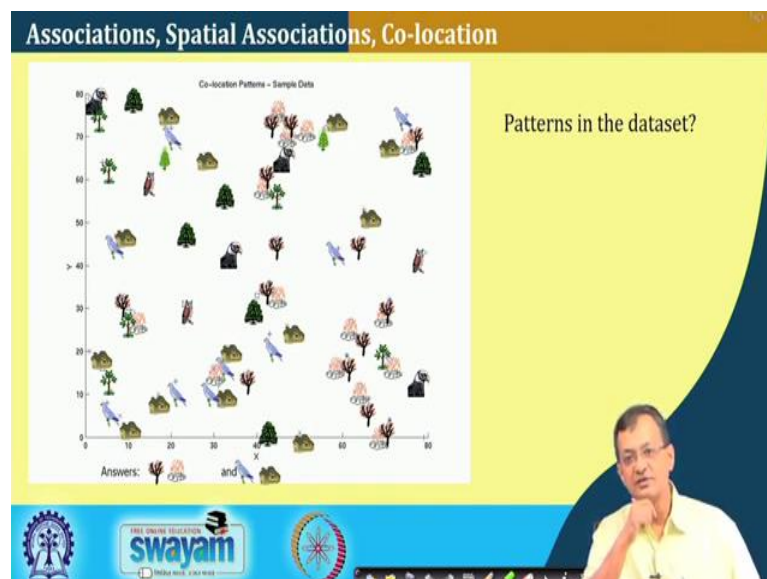
So, how we can find out this range? And this is widely association rule mining is widely used for our different business strategies and different market policies and market basket policies and things. So, how we have a huge item set? Then try to find out what are the interestingness measures into the things which influence here.

So, classical method, association rule given item times and transaction, assume spatial data can be decomposed into transaction, however, each decomposition may alter spatial patterns, right. Spatial methods, spatial association rules, spatial autocorrelation. So, association rule or co-location rules are first filter to reduce the number of pairs of rigorous statistical analysis that is correlation analysis, cross-K-function and spatial interaction etcetera.

So that means, it basically bring down the search space, right, so to have a things. So, but any association rule things what we require is a, important things what we require is more of a how much support and confidence are we are having. Support what it says we will see within couple of site

Support what we tries to say that this feature set that is the feature set a union b, a implies b if I say how frequently occurs in the total my data set, right. And the confidence given those things I want to see that if a occurred b has occurred how many cases. So, I should have a minimum I may have a minimum support and confidence level which tells that what is the support confidence level, how strong is my association rule.

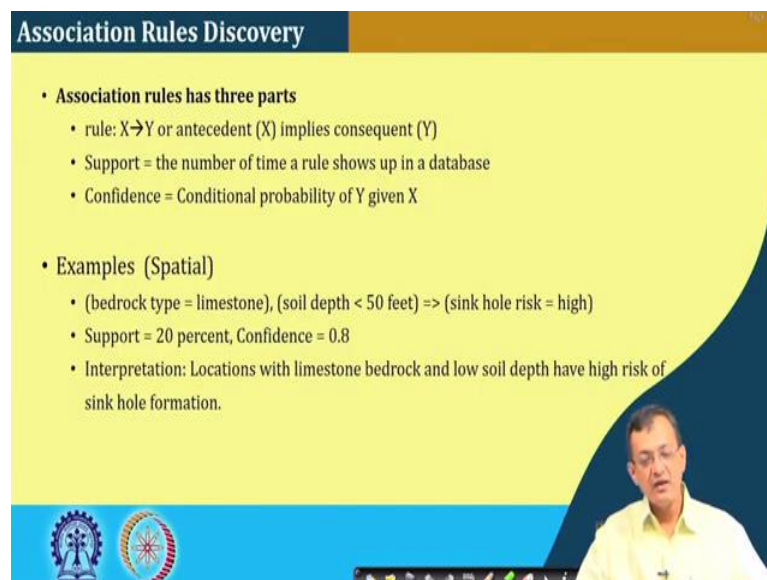
(Refer Slide Time: 13:52)



Like here one thing which we will find in that Shashishekhar's book that co-location patterns. Now, looking at this I may not find apparently any pattern or you may have to look deep into the things; however, we can see that whether there is a pattern we see that these features are occurring more frequently or very frequently whenever there is this part. And there is some sort of a this hut types of things are there or whenever these two things are occurring together or the frequency of occurring is there.

So, I can say there is a pattern in this data set whenever this occurs and this occurs with this amount of confidence we can say. Once we can have we can go little deep into the things by having more looking more into more statistical methods and see that how good or how strong is the association or correlation between them.

(Refer Slide Time: 14:48)



Association Rules Discovery

- Association rules has three parts
 - rule: $X \rightarrow Y$ or antecedent (X) implies consequent (Y)
 - Support = the number of time a rule shows up in a database
 - Confidence = Conditional probability of Y given X
- Examples (Spatial)
 - (bedrock type = limestone), (soil depth < 50 feet) \Rightarrow (sink hole risk = high)
 - Support = 20 percent, Confidence = 0.8
 - Interpretation: Locations with limestone bedrock and low soil depth have high risk of sink hole formation.

The slide features a yellow background with a blue header and footer. A video inset in the bottom right corner shows a man in a light blue shirt speaking. Logos of institutions are visible in the bottom left corner.

So, association rule discovery as many of must be knowing that rule X implies Y or antecedent X implies the consequent Y, right. There is a support, the number of times a rule shows up in a database, right that or the item set in X and Y that overall item set in the database confidence conditional probability of Y given X is a occurred, right. So, conditional probability Y given X is occurred.

Like there is some example like bedrock type equal to limestone soil depth is less than 50 equal to sinkhole risk is high, right. So, if this is the situation that sinkhole risk is high. The support is 20 percent. So, all the out of all data set 20 percent of the cases these things has reflected, whereas, if this happens this left hand side happens, this will happen

the confidence is 80 percent or 0.8, right. So, both are important. So, whenever we want to find out we may look into for the minimum support level and given that support level what is how much confidence I am having on the things.

Interpretation, location with limestone bedrock and low soil depth have high risk of sinkhole formation, right. So, this may be the interpretation, location of a limestone bedrock and low soil depth has a high risk of sinkhole formation, if that support is acceptable. To me that is, but at least this confidence is pretty high if that is there, there is a risk is there.

(Refer Slide Time: 16:24)

Association Rules: Formal Definitions

- Consider a set of items, $I = \{i_1, \dots, i_k\}$
- Consider a set of transactions $T = \{t_1, \dots, t_n\}$
 - where each t_i is a subset of I .
- Support of C $\sigma(C) = \{t \mid t \in T, C \subset t\}$
- Then $i_1 \rightarrow i_2$ iff
 - Support: occurs in at least s percent of the transactions: $\frac{\sigma(i_1 \wedge i_2)}{|T|}$
 - Confidence: Atleast $c\%$ $\frac{\sigma(i_1 \wedge i_2)}{\sigma(i_1)}$

So, there are standard measures. So, we find that I as the item set i_1 to i_k , and consider a set of transaction that number of events recorded in case of a non-spatial or department does not is it, it may be there how much how much things has been bought at together. I want to find out that if a if a customer buys this, this item or this set of items he or she will buy that set of item. So, this work together, where each t_i is a subset of item set, right. All this transaction is one of those items, right. So, that feature set what we are looking for.

So, support we are telling that if t belongs to this transaction set, so this t is a subset of this a transaction set; that means, how much percentage of the transaction contains that particular set C . And then i_1 implies i_2 or for that matter set of i_1 implies as

another set of i things item set implies to item set, then we want to find out that how much they occurred together and in total where this i has occurred.

In other sense, if i has occurred how frequently that i 2 has occurred or if I say if i 1, i 2, i 1, i 3, i 5 has occurred how frequently i 2 and i 4 has occurred or i 2 or i 10 has occurred so, this gives me a this association strength of this association of the things. So, if it is highly occurring thing and I have a good confidence, so at the better support and confidence I say this is likely to occur. Then I can have more finer statistical methods to look into the stuff.

(Refer Slide Time: 18:14)

Spatial Association Rules

- Spatial Association Rules
 - A special reference spatial feature
 - Transactions are defined around instance of special spatial feature
 - Item-types = spatial predicates

So, there are association a spatial reference to the spatial features, right. Transaction are defined around instance of the spatial features, things those here the what do you means by what is the transaction is that some instances of the special spatial feature set and item set is the spatial predicates, right. So, I have defined like we have seen limestone and other things and going to the sinking. So, these are my item set so to say, right. Like, bedrock type limestone, soil, low soil depth and these are my and risk of sinkhole or sinkhole risk these are my item set. So, these are all spatial features.

(Refer Slide Time: 18:58)

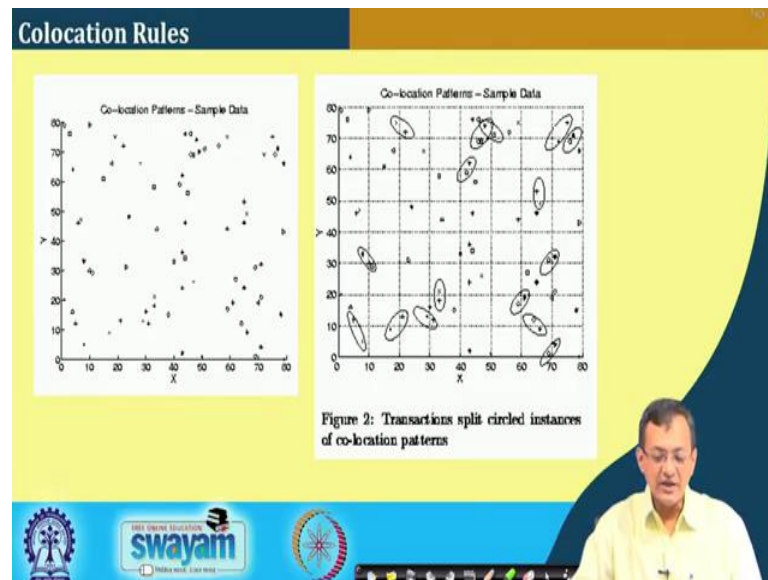
Co-location Rules

- Motivation
 - Association rules need transactions (subsets of instance of item-types)
 - Spatial data is continuous
 - Decomposing spatial data into transactions may alter patterns
- Co-location Rules
 - For point data in space
 - Does not need transaction, works directly with continuous space
 - Use neighborhood definition and spatial joins
 - “Natural approach”

So, another means related thing is the co-location rules. So, what we say there? That thing that association rules transaction subsets of the instances of the item sets, right how much things transactions are there. Spatial data is continuous, right. Decomposition spatial data into transactions alter the pattern, so it is a continuous phenomena. So, if you block this transaction, this transaction it is a difficult or it may break the pattern. I may lose the pattern by breaking them into the things, right.

So, co-location from the point data in space does not need transaction works directly on the continuous space. So, this co-location it works on the directly on the things, use neighbourhood definition and the spatial joints. So, it is more of a quote unquote natural approach for spatial data analysis, right. So, this sort of co-location data mining.

(Refer Slide Time: 19:57)



So, I can say that if it is a things again this figure taken from Prof Shashishekhar's book. So, co-location pattern sample data set and then we try to find out that what are the things. So, transaction split circled instances on the co-location patterns. So, these are the patterns which are co-located patterns, right. See, they to they occurred together a number of frequency. Then I have to find out that how frequently they are occurring together and if it is over my expected support and confidence then I accept that that is a.

(Refer Slide Time: 20:31)

Co-location rules vs. Association rules

	Association rules	Co-location rules
Underlying space	discrete sets	continuous space
item-types	item-types	events /Boolean spatial features
collection	Transaction (T)	Neighborhood (N)
prevalence measure	support	participation index
conditional probability metric	$\Pr[A \text{ in } T \mid B \text{ in } T]$	$\Pr[A \text{ in } N(L) \mid B \text{ at location } L]$

Participation index = $\min\{\Pr(f_i, c)\}$
 Where $\Pr(f_i, c)$ of feature f_i in co-location $c = \{f_1, f_2, \dots, f_k\}$:
 = fraction of instances of f_i with feature $\{f_1, \dots, f_{i-1}, f_{i+1}, \dots, f_k\}$ nearby
 $N(L)$ = neighborhood of location L

So, if we try to have a some sort of a I should not say comparative study, but see that where they feeds, like for association rule and co-location rules like underlining space is discrete, this is continuous, items say item types is item sets here events, Boolean spatial features whether there or not. Collection is a transaction set of collection, here more of a neighbourhood, right. So, it is not that how much occurred, but it is more important that why which neighbourhood that is there.

Prevalence measure say the support and followed by confidence, where here the participation index like how much participation is there. And here the probability we want to find out A in T, means conditional probability and where here we want to find out that with respect to the neighbourhood calculation, how many within this neighbourhood and out of that within the neighbourhood how is the how many occurrence of B.

So, participation index is we calculate by finding those minimum of the probability of f_i and c , where f_i and c are the measure of feature of feature f_i on co-location c that these are the co-located feature sets. Fraction of the instances of f_i with feature f_1, f_2, f_k nearby, so N_L is the neighbourhood location or in other sense within a neighbourhood location we want to find out how many such pieces sets are occurring together. So, that is our way of looking at the co-location thing.

And it makes more lot of sense because we want to see some of the co-located events. Some of the events are co-related events, right they may be in different space, but this happens that will happen maybe in a different temporal scales. So, these are co-located or correlated events or something co-occurring events, some things which occurs at the same time. So, they may be in the neighbourhood which is more not that may not be a good ah what we say interesting feature, because definitely if it is raining in IIT, Kharagpur in residential area it is expected there is raining in the academic area also, right.

But if it is raining here and co-occurrence in say 50 kilometers away also it is raining then it may be a co-occurred events. So, if it is frequently happening. So, that is a phenomena if I say that this happens that will happen that may be lot of underlining metrological physics phenomena into this. What we are looking as more as the data driven approach to this problem.

(Refer Slide Time: 23:10)

Clustering

- **Clustering**
 - process of discovering groups in large databases.
 - Spatial view: rows in a database = points in a multi-dimensional space
 - Visualization may reveal interesting groups
- A diverse family of techniques based on available group descriptions
- Example: Census
 - Attribute based groups
 - Homogeneous groups, e.g. urban core, suburbs, rural
 - Central places or major population centers
 - Hierarchical groups: NE corridor, Metropolitan area, major cities, neighborhoods
 - Areas with unusually high population growth/decline
 - Purpose based groups, e.g. segment population by consumer behaviour
 - Data driven grouping with little a priori description of groups
 - Many different ways of grouping using age, income, spending, ethnicity, ...

So, next what we see is that clustering. So, process of discovering groups in large databases, spatial view, rows in database, points in the multi-dimensional space, right in a spatial view rows in a database the points in a multi-dimensional space, right, it is a distributed. Visualization may reveal interesting groups. So, only looking at the data set may not be that, having visualization may have a interesting feature set.

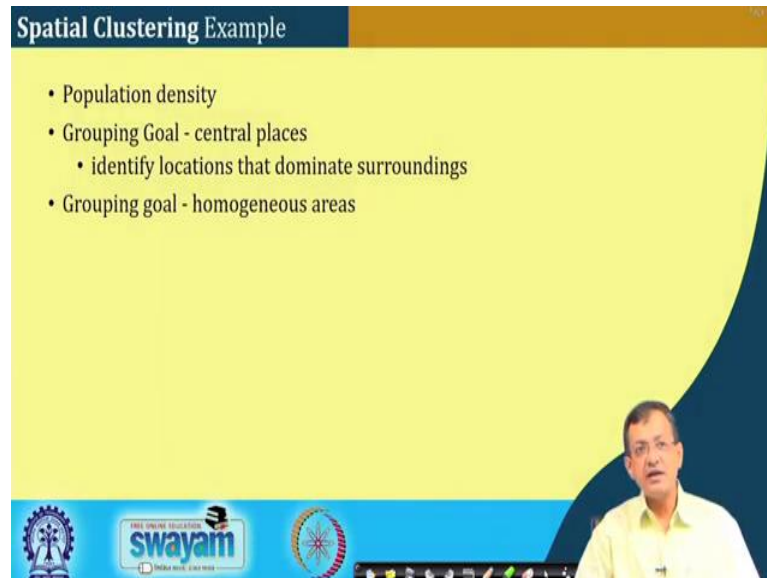
Like a diverse family of techniques, a diverse family of techniques based on available groups and description. So, example we have a census data attribute based groups, homogeneous group, urban core, suburbs, rural and type of things based on the census data. Central place or major population centers. Hierarchical group, northeast corridor, metropolitan area, major cities neighbourhoods. Areas with and usually unusually high population with growth and growing or declining. So, these are the things which can be manifested when we have a spatial manifestation of the things, right.

And purpose of purpose bases group that the segment of population and consumer behaviour. Data driven groups of little prior like this sort of population is more driven toward this sort of product or this sort of activities and type of things. So, different, many different ways to grouping like we can group by age, income, spending etcetera type of things, right.

So, some are, some of the things are spatial some of the things are non-spatial in nature, right, but when we look at the spatial spread on a visualize over a map that keeps us

interesting feature. So, what we are looking at the clustering more on the clustering based on the your own the spatial context.

(Refer Slide Time: 25:06)



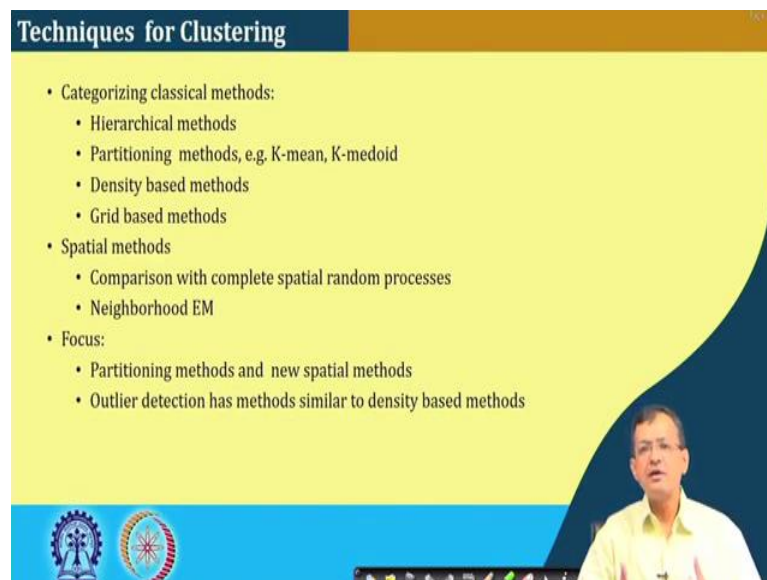
Spatial Clustering Example

- Population density
- Grouping Goal - central places
 - identify locations that dominate surroundings
- Grouping goal - homogeneous areas

The slide features a yellow background with a dark blue curved border on the right. At the bottom, there is a blue banner with logos for 'swayam' and other educational institutions. A presenter is visible in the bottom right corner.

That there are several clustering methods which we use on the spatial.

(Refer Slide Time: 25:14)



Techniques for Clustering

- Categorizing classical methods:
 - Hierarchical methods
 - Partitioning methods, e.g. K-mean, K-medoid
 - Density based methods
 - Grid based methods
- Spatial methods
 - Comparison with complete spatial random processes
 - Neighborhood EM
- Focus:
 - Partitioning methods and new spatial methods
 - Outlier detection has methods similar to density based methods

The slide features a yellow background with a dark blue curved border on the right. At the bottom, there is a blue banner with logos for 'swayam' and other educational institutions. A presenter is visible in the bottom right corner.

And as you find that clustering is a standard technique for all mostly the datas, several data driven for our non-spatial data and also for image video processing type of things. And those will also leverage on those techniques while we talk about the clustering.

So, there are several techniques for clustering what we see that like hierarchical methods, partitioning methods, K-mean, K-medoid and different those are very much here, very much in used in the spatial context.

Additionally, for spatial methods though we need to consider that comparison between the complete spatial random processes, right. So, and also the phenomena of the neighbourhood, right. So, as there are prevalence of autocorrelation things. So, the phenomena of neighbourhood. So, what we here it is the more focus is on the partitioning methods and new that how to partition the space in a spatial context and outlier detection has methods similar to the density based methods and so and so forth, right.

(Refer Slide Time: 26:16)

Outliers

- What is an outlier?
 - Observations inconsistent with rest of the dataset
 - Techniques for global outliers
 - Statistical tests based on membership in a distribution
 - $\Pr[\text{item in population}]$ is low
 - Non-statistical tests based on distance, nearest neighbors, convex hull, etc.
- Spatial outliers?
 - Observations inconsistent with their neighborhoods
 - A local instability or discontinuity
- Techniques for spatial outliers
 - Graphical - Variogram cloud, Moran scatterplot
 - Algebraic - Scatterplot, $Z(S(x))$

So, now what is the outlier? As you must be knowing that is a observation inconsistent with the rest of the data sets. So, that you have a data set and the observation may be inconsistent with the rest of the data set. So, that is a problem. So, techniques for global outlier. So, statistical test based on members in a distribution that there are statistical test in noting the probability of item in the population is low and type of things.

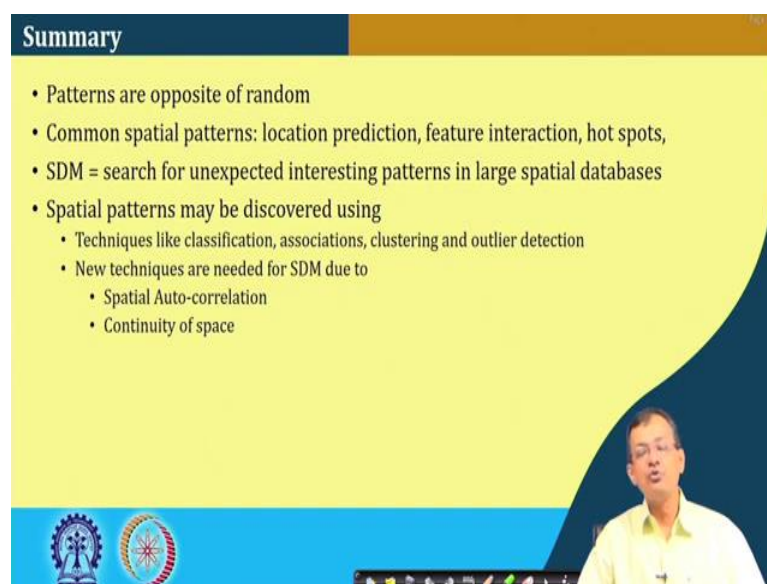
Non-statistical test based on distances, nearest neighbour, convex hull and type of things. And there may be other different semantics consideration in case of a spatial data like if I find a phenomena which is not likely to occur in this particular things that may I consider as a outlier, so there are different statistical method, there are other things.

So, for spatial outlier observers in inconsistency with the neighbourhood. I suddenly find this temperature of IIT, Kharagpur one region is say recorded one sensor is recording the temperature 1.5 times that the next sensor of the which is in the next say electric pole or something, right. So, that is the outlier. It is unlikely to be in this neighbourhood that thing will be there, right. So, a local instability or discontinuity can be a things like that, right. So, techniques for spatial outlier. Graphical - variogram cloud, Moran's scatterplot. Algebraic, scatter plot and type of things. These are the different standard techniques for spatial outliers which are being used.

So, what we try to discuss in this today or rather in the spatial analysis context is more of how do I use the data into having meaningful or interesting feature out of it or extracting so called quote unquote knowledge from the huge volume of data set. As the spatial data set has a spatial, spatial feature of it is on a space time scale it is spread, not only that there is a spatial feature of autocorrelation. They are not un, that sample set are not totally independent, right.

So, like temperature of a region is related to the humidity related to the things, say congestion level of a region is read is not that independent of the next region, right. So, the congestion level they are related. So, this is this type of things need to be captured in these things.

(Refer Slide Time: 28:37)



Summary

- Patterns are opposite of random
- Common spatial patterns: location prediction, feature interaction, hot spots,
- SDM = search for unexpected interesting patterns in large spatial databases
- Spatial patterns may be discovered using
 - Techniques like classification, associations, clustering and outlier detection
 - New techniques are needed for SDM due to
 - Spatial Auto-correlation
 - Continuity of space

So, what in this particular lecture; we try to look at the patterns are opposite to randomness. Common spatial patterns, location prediction, feature interaction, hotspots, these are the some of the common spatial patterns. Spatial data mining starts for unexpected interesting patterns in the large spatial databases. Spatial patterns may discovered using techniques like classification, association rule mining, clustering, outlier detection.

There are various new techniques which are, I should not say new techniques it is the techniques which are more for the spatial data mining type data mining or spatial data analysis is the spatial autocorrelation and contiguous of space. So, with this let us conclude our discussion on this particular topic. And we will continue our discussion in our subsequent talks.

Thank you.