

Spatial Statistics and Spatial Econometrics
Prof. Gaurav Arora
Department of Social Science and Humanities
Indraprastha Institute of Information Technology, Delhi

Lecture - 06
Spatial Statistics: measures of variation; spatial random function

Hello everyone, my name is Gaurav Arora and welcome to the 5th lecture of Spatial Statistics and Spatial Econometrics.

(Refer Slide Time: 00:30)

The role of Spatial Statistics

- Statistics traditionally is a science of uncertainty or disorder
 - characterizes/models order in disorder
- Spatial Statistics is a science of uncertainty of "spatial nature".
 - characterize order in disorder through space explicitly account for space.

Usually, the first step is to MEASURE spatial disorder

Statisticians } - VARIANCE
 } - Inter-quartile range

Engineers and numerical } - Entropy



In this lecture we will move on to the formal components of spatial statistical modeling, I want to begin with you know articulating the role of spatial statistics. So, statistics traditionally, is understood traditionally, is understood to be a science of uncertainty or disorder.

So, we view the world, the physical world, the social world, the natural world whichever you are interested in as statisticians, econometricians, we view the world as a sequence of random events or random realizations right. So, what we are committing to or declaring when we start to study the real world is that, the real world is fundamentally disorderly or there is fundamental uncertainty attached to the events or occurrences that we observe in the world that we are interested to study right.

And statistics is a science that allows us to systematically understand or study, you know, this uncertainty or disorder right. So, what statistics does is that it characterizes or models “order in disorder.” So, it tries to figure out the systemic features some things that we can explain right. So, there is a lot of uncertainty lot of randomness, but some of it a part of it is explainable, which is orderly and that order is what we are out here to figure out, to figure right; so, to discover.

Spatial statistics, as we have discussed is an extension right. Spatial statistics on the other hand is a science of uncertainty of *spatial* nature, ok. As we had said earlier that spatial statistics will add this dimension of location or “where” to the statistics that we understand on the intensity of events, the count of events, the frequency of events and so on and so forth. So, we are not only going to stop at learning what and how, we are also going to start learning the systemic features of where in space are these uncertain events happening.

So; obviously, through that extension we can say that spatial statistics will characterize order in disorder through space right through space means it will explicitly account for space, ok. So, this is a fundamental difference between what statistics is understood to be doing traditionally some of you are coming from the statistics background, you would have understood the first component.

The second component is something that you will learn as a fresh understanding of the spatial nature or exploiting space to figure out systemic features of a disorderly world. Those of you who do not come from statistics background we will provide you a basic understanding of you know how traditional statistics and spatial statistics will differ technically in a little bit, right.

So, usually the first step the first step in spatial analysis is to measure spatial disorder right. So, measuring this spatial disorder would mean that you are you know somehow trying to measure variation or the type the variety of you know locational features or spatial features I can describe in the data.

So, the tools that are available to us are variance, which is quite natural to understand. The second tool is called as the inter-quartile range if you have not heard of it do not worry we will come back and explain it in a bit. The third is called as entropy, ok. These are three most popular fundamental of you know metrics of measuring disorder.

So, I am trying to find order in disorder as a spatial statistician, but first I need to be able to quantify disorder I need to be able to understand what is the level of variation in my data to begin with what I am I out to explain what is the problem I am out to address right.

So, that is the aim that I am trying to you know achieve right. So, variance and inter-quartile ranges these are popularly used or you know extensively sort of applied or employed by statisticians, econometrician and so on as biostatisticians and so on and so forth.

Entropy on the other hand is mostly used by engineers and physical scientists. So, if you come from physical sciences, you know background perhaps you must have heard of entropy. What we are learning here is that entropy does a similar job to what variance does and this is something we will again see in a little bit going forward ok. An entropy provides what is called as the philosophy of technology.

So, this is something we will spend a little bit of time on because this is interesting and it also sort of allows us to model certain really important spatial features in the real world especially regional science or urban science right. So, we will come back to both these categories of measuring disorder you know in space.

(Refer Slide Time: 07:46)

- Second step is to describe/summarize spatial patterns, spatial dependence and/or spatial auto-correlation.
 - Summary statistics for spatial data
 - Spatial models for variogram estimation
- Third, explain spatial processes underlying the spatial patterns that were discovered in step 2.
 - Principles of econometrics
 - Correlation to causation. Why?



The second step the second step is. So, the first step was to describe measure disorder of the problem that we are interested in and the second step is to describe or summarize spatial patterns, spatial dependence and I am going to say and or spatial auto correlation in data.

So, the first step was that I am trying to understand what is the extent of uncertainty, in the second step now I am moving slowly I have taken one step to figure out some kind of order in this disorder. So, I am trying to do a pattern recognition exercise I am trying to summarize what are the patterns do I see here do I see a trend from east to west from north south sorry northwest to northeast or northwest to southeast and so on and so forth right.

So, I am now trying to figure out what are the patterns, do I see clusters, some kind of networked clusters and some corners or some regions of the data where things might be moving similarly than other regions right. So, here we will study summary statistics for spatial data right.

So, of course, you know whenever we look at any data set, we start to do we start to look at the summary statistics you know that is mean, variance, standard deviation, coefficient of variation different percentiles, you know the min value the max value the value at the 75th percentile, the 25th percentile the median and so on and so forth right.

So, we will basically articulate these tools or these metrics also for spatial data. Then we will look at something really unique to spatial data that are called as the variogram devices. So, we will say spatial models for variogram estimation. So, the variogram is quite typical to spatial analysis. It is a metric of spatial dependence of spatial autocorrelation, it is an analog of a correlation device that you are used to in summarizing you know traditional statistical data sets right.

So, variogram is a measure of spatial dependence and we will be, you know studying this particular tool in detail which will be a discrete sort of you know extension in terms of your understanding of you know applied statistics till now ok. Third step in this exercise of studying or analyzing spatial data is to explain spatial processes underlying the spatial patterns that were discovered in stage 2 or step 2.

Here we employ principles of econometrics principles of econometrics and we traverse the journey of correlation to causation and you know we go from how, to what, to where, to now, why, why do we see what we see in space? Why do we see a west to east gradient in real estate prices in the national capital territory of Delhi? Why do we not see reverse? Why do we not see a gradient from south to north? Why do we not see a gradient from southwest to north east right?

Why do we see the gradient the way we see it? Why do we see the patterns the way we see them right. So, we are trying to answer *why* this is a fundamental distinction from you know statistics to econometrics and that is why this course is called as spatial statistics and spatial econometrics due to this third component, which involves certain advanced, you know tools to study spatial data right.

Now, before we move on to talk about the first step which is measuring order and disorder which is the variance, the inter-quartile range and the entropy, we will first you know try and understand what is a spatial random variable how do we articulate randomness in space right.

(Refer Slide Time: 13:05)

(PDF) Probability Distribution function } in spatial domain / as a function of location
 (CDF) Cumulative Distribution function }

Traditional Stats
Spatial Stats

Random variable
 $X \sim f(X)$
 $x \rightarrow$ a realization of X

So, we are going to first study probability distribution functions and cumulative distribution functions, distribution function, in spatial domains or you can say as a function of location. How does the idea of a PDF or a CDF which are very primitive you know, I expect that most of you would understand what a PDF or CDF is if you do not that is fine I mean you can go back and study these are basic probability and statistics you know entities that you can find in any you know basic undergraduate course of probability and statistics, right. So, let us make a distinction let us start with making a distinction between traditional statistics, so, we will on the left hand side you will articulate what happens in traditional stats and on the right hand side I will articulate what happens in spatial stats.

So, in traditional statistics we are used to we are used to starting with a random variable. So, say you have a random variable X which is distributed with probability distribution function

$f(X)$ right and you have a realization of X which we denote as x which is a realization of X , random variable X . So, when we say when I say that X has a probability distribution function X follows a probability distribution function what I am really saying is that X can take a range of values it can take multiple values with different probabilities or a different frequencies.

So, how do we understand that, well you know we can plot on the x-axis we can plot the random variable X all the different values on x-axis are the possibilities that X can take you know, on the real number line, which is on the positive side of the real number line as I have drawn it although we do not have to we do not have to respect that restriction all the time right. On the y-axis I have what is called as the frequency of observing different values of you know possible values of this capital variable X right.

So, you can imagine a histogram right. So, histogram gives you a frequency of different values on the x-axis you know what will be the frequency of smaller values smaller you know values of this random variable X what would be the density of very large values of X and then what would be a density of you know some values in between.

But whatever we do you know we are sort of used to looking at a function which looks as nice as a bell curve, which tells me that there is very low density with which or frequency with which I am going to observe the extreme low values, the extreme high values and there is going to be quite a high density with which I expect to observe you the middle values or the intermediate values of X right.

This is not necessarily the only way we can describe a probability distribution function of X right. We can also define or describe the PDF of X as highly skewed as on your screens which basically says that there is very high frequency of low values of X and as we go on to the higher values the frequency declines pretty fast and becomes almost near to zero pretty quickly right. This is a different shape this is a different description of you know of the frequency distribution of X .

But what is common is that X can take multiple values a range of values and to each value is attached a probability distribution function ok. So, that is the idea of a random variable and if we want to understand you know what is a realization. So, where would this small x would be? Well, it would be you know some value on this x-axis that we have you know that we have written down.

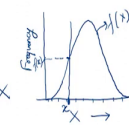
So, this particular realization that would have happened with this level of frequency that I am marking on the y-axis f of small x is a realization of X which we usually get to observe in the real world. So, what we get to see is the realization, what we want to understand is the probability distribution. So, that is the complex pathway that we want to traverse and hence you know we want to sort of you know we want to sort of employ certain tools of statistics ok, what happens in case of spatial statistics.

(Refer Slide Time: 19:11)

PDF } Probability Distribution function } in spatial domain / as a function of location
 CDF } Cumulative Distribution function }

Traditional Stats

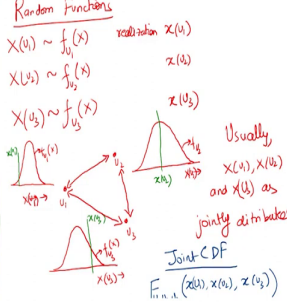
Random variable
 $X(u) \sim f_u(X)$
 $x(u) \rightarrow$ a realization of X



Spatial Stats


Random Functions

$X(u_1) \sim f_{u_1}(X)$ realization $x(u_1)$
 $X(u_2) \sim f_{u_2}(X)$ $x(u_2)$
 $X(u_3) \sim f_{u_3}(X)$ $x(u_3)$



Usually,
 $X(u_1), X(u_2)$
and $X(u_3)$ as
jointly distributed

Joint-CDF
 $F_{X(u_1), X(u_2), X(u_3)}$



So, first of all because there is this locational understanding, the first thing first that is going to happen is I am going to have a location appended to each you know entity that I am describing in this random you know universe right. So, X which is a random variable it is defined with an index u which is a location in space the density f is also described or is also you know tied to a given location.

And then the realization that I am talking about is also tied to a given location right, but in case of spatial statistics instead of random variables we are usually working with what are called as random functions. So, how is a random function different from a random variable? So, first up you can have data for different locations right we typically will not have data for just one location we are going to have data for multiple locations.

So, let us start there. So, we are going to have a random variable X at location u_1 which is going to have a PDF $f_{u_1}(X)$. You are going to have random variable X at location u_2 which is going to have a PDF $f_{u_2}(X)$ and let us say there is a third location where I observe this random

variable X with density $f_{u_3}(X)$ and I am going to have you know realizations corresponding to these as small $x(u_1)$ small $x(u_2)$ and small $x(u_3)$.

So, I am going to now give you a visual representation of random functions. So, the first thing that we realize here is that we have three locations u_1 , u_2 and u_3 . So, let us go ahead and draw these three locations. So, I have u_1 , u_2 , and u_3 sorry about that. So, u_1 , u_2 and u_3 at each of these locations I have a random variable X of u_i which will have a PDF. So, I can draw that PDF now for u_1 right.

So, I have a $f_{u_1}(x)$ right. So, ok $f_{u_1}(x)$ at location 2, I am going to have $f_{u_2}(x)$ and then at location 3 I am going to have $f_{u_3}(x)$ right. So, f at u_3 rather right and I am on x-axis on all these graphs I have the random variable X defined for different locations ok and I have realizations. So, first of all I could have different shapes of density functions at the three locations second of all I might have a realization from different regions of these you know density functions.

So, for example, for the first location I might draw my first realization might come from the lower region of x_{u_1} . So, I might have a x small $x(u_1)$ on the left side of the distribution for u_2 it might come from somewhere around the expectation or the mean of that distribution.

So, I am going to have $x(u_2)$ you know ah delineated like you have in front of your screen and then $x(u_3)$ is going to be you know let us say coming from the right region or the higher side of the probability you know of the higher values you know from some higher values of $x(u_3)$ right. So, the right hand side of the PDF of 3 and I have these three you know locations. So, if I have locations when I do spatial statistics the first thing I must do it is I must be able to articulate distance.

So, if I am working with Euclidean spaces you know I am going to have distances basically defined as the L2 norm as we have studied earlier. So, you have three data points which are at given distances separated by given distances in space and at each you know data point at each location you are going to observe a look realization that is that belongs to a random variable which can be distributed by a distinct PDF.

Now what differentiates you know random functions from random variables? Is that usually we have the you know $x(u_1)$, $x(u_2)$ and $x(u_3)$ as jointly distributed ok. Mathematically joint distributions would mean that when I try to explain what value will be observed at location u_1

it will give me some information about what type of realization might be observed at u_2 and u_3 ok.

So, there is a jointness in the way these three values are realized. So, if you think about the example of groundwater levels right we are we have groundwater you know sitting beneath the surface of the earth, we have dug three wells at u_1, u_2, u_3 when I look at the value of the depth of groundwater at a location that is connected with the groundwater depth at location at the at another location say location 2. Why?

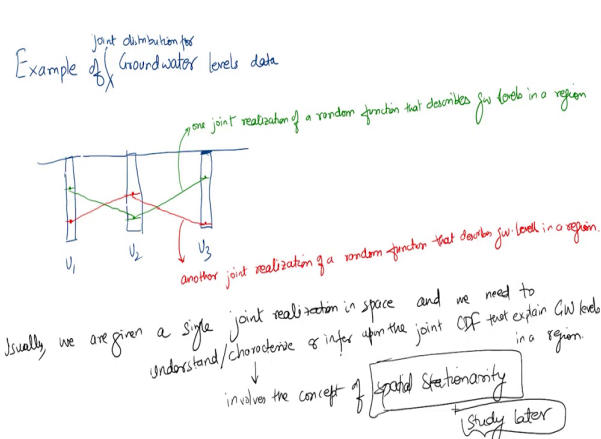
Because water is connected beneath the ground there is continuity “jointness” in the way water levels are going to be evolving beneath the surface in three points which are in proximity in reasonable proximity over a region right.

So, there is going to be this we are if we are going to be able to write a joint CDF right we are going to be able to say F , which has components as locations u_1, u_2 and u_3 jointly understood in a way that when I realize x , small $x(u_1)$ at location 1, small $x(u_2)$ at location 2, and small $x(u_3)$ at location 3 I can write this as the probability measure.

The probability measure that is jointly determined for locations 1, 2, and 3 such that the probability that capital $X(u_1)$ is less than small x that is the realization at location 1. The random variable and the random variable at location 2 is less than its realization at location 2 and the random variable at location 3 is less than its realization at location 3.

This is a joint CDF that characterizes how three values will be determined to better understand let us go back to our example of groundwater levels. So, let us look at an example of groundwater water levels rather an example of joint distribution for groundwater levels data.

(Refer Slide Time: 27:24)



So, here you know on ground what I have is three wells. So, I have three wells where I am able to draw data right monitor data from sorry not draw, but monitor what is the depth of data from ground. So, let us say I have a ground level and I have dug these wells beneath the ground at locations u_1 , u_2 and u_3 right.

And on a given day I have these realizations where I see that groundwater level is at this red dot level at location u_1 , it is at a the red dot level at location 2, which is slightly higher than location 1 and then it is quite you know below the ground quite a bit below the ground at location 3.

The point is that this realization at each individual well is connected with every other realization of these wells that are representing groundwater which is moving in continuum in you know in space beneath the surface of the earth right. So, what is happening is that the three observations are connected with each other. So, when we look at a realization we look at them as triplets right.

So, we do not look at them really as individual realizations, but they are a triplet realization of the joint density of groundwater levels beneath the ground. On another day you can have, you know, groundwater levels different at different levels at these three locations then, your observation will be another triplet which are joined by the green dots or green levels you know levels are demarcated with the green color at these locations right.

So, green is one joint realization of a random function that describes ground water levels in a region in a region with three wells right we could have 300 wells, the x the definition will simply extend itself you know naturally right and the reds are an alternative you know I am just going to say another joint realization of a random function that describes groundwater levels in a region ok.

Now, the thing is that usually we are given a single joint realization in space realization in space and we have to then infer upon the joint distribution right and we need to understand or infer upon, understand, characterize or infer upon the joint CDF that explains you know groundwater levels in a region ok. So, that explains the stochastic nature of groundwater levels in a region. So, once you are able to sort of understand what is the parent CDF you can then you know start to model things much in much more detail ok.

So, random functions are really jointly distributed random variables. So, there are multiple random variables ok. So, we are talking about multivariate you know variables multivariate distribution something that we a multivariate is a term that was also used at the beginning of this lecture right. So, we are talking about multivariate you know a distribution. So, you have more than one variable these variables are delineated by location.

So, the multiplicity is derived from location and hence the delineation of space in conducting statistical analysis right. So, the trouble is that I have one realization and I must tell you what is the parent distribution. So, I must be able to predict if you were to go ahead and take another realization what should it look like in the sense how it is dependent and what kind of mean values you are going to look at and so on and so forth.

This you know trajectory from using one realization to providing a characterization of the entire distribution, joint distribution is complicated and it involves this concept involves the concept of “spatial stationarity”. This is one of the most important concepts that we will study later ok. We will study this concept later in the course, but spatial stationarity is the key assumption really ok.

It is something that we must go in with we have to be confident that our you know we have spatial stationarity to be able to you know expand our understanding of the world from a single realization to what might be happening in terms of joint distribution in space right so, that is that so. So, now, you know as a next step what we are going to do is we are going to define variance.

(Refer Slide Time: 34:20)

Define Variance, standard deviation, inter-quartile range, Entropy



Given a sequence of realizations of a r.v. $X: x_1, x_2, \dots, x_N$

$$\text{Var}(X) = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}$$

$$\text{S.D.}(X) = \sqrt{\text{Var}(X)}$$

Interquartile Range $\in [Q_1, Q_3]$

Organise your data in ascending order

Let's say $x_1 < x_2 < x_3 < x_4 < \dots < x_N$

$Q_1 = x_{25}; Q_3 = x_{75}$

IQR $\in [x_{25}, x_{75}]$

Next lecture
Probability



So, we are going to go back to our step 1, now we have understood random functions that is what we are going to use to characterize spatial statistics. So, we want to first measure disorder. So, we had a variance, we had you know we can say standard deviation although it is just it is just a simple extension of variance, it is just a square root of variance.

Then you have what we said was inter-quartile range and finally, we had what we said what we called entropy right. So, given a sequence of realizations of a random variable X ok I am not invoking space now I am first defining these metrics just I am just making a quick we are just doing a quick recall so that you can get it and you can we can then you know apply these metrics to spatial data. So, I have a sequence of realizations of a random variable X let us say x_1, x_2 and all the way to x_N .

So, I have capital N realizations, then variance of X is written as:

$$\text{Var}[X] = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}$$

The standard deviation of X is:

$$\text{sd}[X] = \sqrt{\text{Var}[X]}$$

and the inter-quartile range inter-quartile range a little bit interesting and I will talk about why do we care about inter-quartile range? Why do we even study it? Well, the inter-quartile range is an interval that lies between the first quartile and the third quartile of data.

So, what you do is you order your data order your data in ascending sorry about that just ok in ascending order ok. So, you organize your data, you organize your data in ascending order of values right. So, organize your data right. So, you know let us say you know you have a lucky situation you know let us say we have x_1 is less than x_2 is less than x_3 and less than x_4 and so on.

So, the data are already ordered as presented in the sequence and you know once you have the data organized in that way you can go to the 25th percentile that is 25 percent of the data one fourth of data values are smaller than particular that particular value, let us say you know that value is x_{25} right and the that is called the first that is called as the first quartile the third quartile which is a 75th percentile, which basically means 75 percentage of my data values are smaller than this particular value let us say the third quartile is x_{75} , then the inter-quartile range is between x_{25} and x_{75} .

Formally IQR can be written as this inverse of CDF at 25th percentile and inverse of CDF at 0.75 percentile with closed bounds. So, that is that for this lecture in the next lecture we will study entropy right we will study entropy we will see why do we need entropy we have the variance, we have the standard deviation, we have interquartile range, why do we need entropy right. So, we will go over entropy in more formal sense.

And then finally, how does it help us understand spatial data right and now I want to spend couple of minutes just solving the class exercise that we opened up in the previous lecture. So, you know I had given you three situations and I had asked you what kind of spatial data model would be most appropriate in those situations. Let us go right to that and spend a couple of minutes understand that and end this lecture today.

(Refer Slide Time: 39:12)

Geostatistical model ; Lattice Data ; Point pattern Data

Class Exercise - Identify appropriate spatial data model for the following variables

GS data

Variable to be modeled

Data ID	X- coordinate	Y- coordinate	GW Well Depth
1	78	28	-5
2	79	28	-3
3	78	27	-30
4	79	26	-10
5	82	26	-1
6	85	25	1
7	83	22	-20
8	77	24	-12
9	90	25	-8
10	75	30	-1

I

Lattice Data

Variable to be modeled

Data ID	District	GW Well Depth
1	Bagpat	-5
2	Bahadshahr	-3
3	Badaun	-30
4	Chandauli	-10
5	Chitrakoot	-1
6	Dewari	1
7	Etah	-20
8	Etawah	-12
9	Faizabad	-8
10	Farrukhabad	-1

II

Point pattern Data

Variable to be modeled

Data ID	X- coordinate	Y- coordinate	GW Well Depth	Dry Well?
1	78	28	-5	No
2	79	28	-3	No
3	78	27	NA	Yes
4	79	26	NA	Yes
5	82	26	-1	No
6	85	25	1	No
7	83	22	NA	Yes
8	77	24	NA	Yes
9	90	25	-8	No
10	75	30	-1	No

III



So, here we are this is the exercise from previous lecture towards the end of previous lecture where I had given you a little exercise where you know I wanted you to identify the appropriate spatial data model for following variables. Remember we have three models we have the geostatistical model sorry about that ok.

We have the geostatistical model, geostatistical data, we have the lattice data which are more coarse administrative boundaries remember and finally, we have the point patterns data where the points themselves are random variables.

So, quite clearly the first one is different wells with coordinates in a domain right. So, there are 10 wells 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 right. So, I have 10 wells they are going to be in a domain and I am observing certain ground water level depth now of course, ground water is going to be everywhere I am only able to dig 10 wells and observe values at those wells.

So, this is a clear case where the domain D itself is a continuum right it is a continuum. So, if this first example is the case of the geostatistical data. The second example quite naturally is districts data and districts. So, I can only do as much as these ten districts the shape and form they may come in is not controlled by the analyst that is the statistician. So, I have 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 districts.

So, I am bound by the district boundaries there is some restriction this is a classical case of a lattice data ok. One more way to think about it is that you know once you identify the most

appropriate model think about whether you can apply an alternative model maybe you can right maybe you can maybe you cannot right. So, in the third case you have the you know you have again you have groundwater wells.

So, you have these wells and you are observing rod water level depth, but the variable of interest is whether or not you have a dry well.

(Refer Slide Time: 41:41)

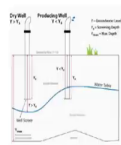
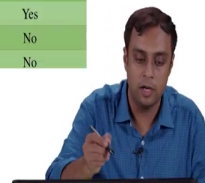


Image courtesy: Saif Ali
Source: Ali and Arora (2021)
[Click here](#) for more details

III

Variable to be modeled

Data ID	X -coordinate	Y -coordinate	GW Well Depth	Dry Well?
1	78	28	-5	No
2	79	28	-3	No
3	78	27	NA	Yes
4	79	26	NA	Yes
5	82	26	-1	No
6	85	25	1	No
7	83	22	NA	Yes
8	77	24	NA	Yes
9	90	25	-8	No
10	75	30	-1	No



And as we saw you know that dry well occurrence of dry well is really dependent on what is the water level beneath the surface and water level going below the level of the well itself is a random process right.

So, then what I am looking at is that whether or not I see a dry well or not is itself a random process hence the 3rd example is a point pattern data. I hope this sort of this characterization helped you better understand what type of data are seen in different settings and I know I hope you are able to apply this knowledge in the future. So, that is it for today.

Thank you for your attention we will meet in the next lecture.

