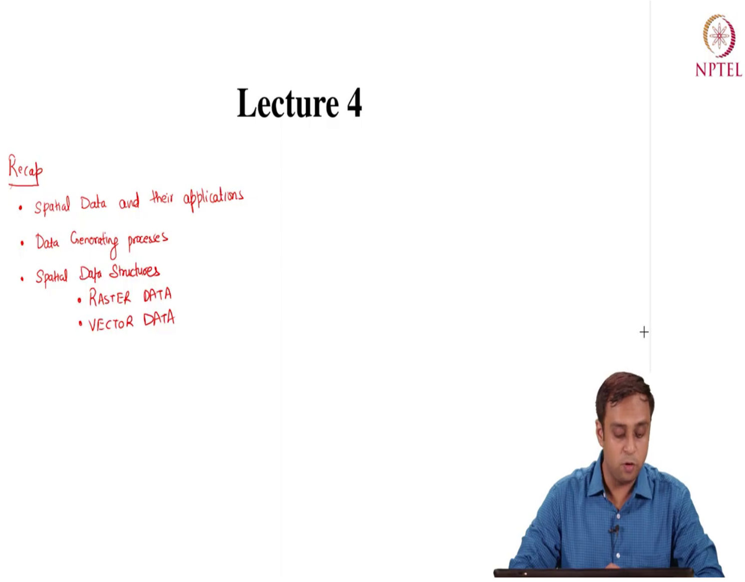
Spatial Statistics and Spatial Econometrics

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

Lecture - 05
A general spatial data model

Hello everyone. My name is Gaurav Arora and I welcome you to the 4th lecture of Spatial Statistics and Spatial Econometrics. So, till now we have you know covered you know we have seen a lot of spatial data and their applications in different domains which gives us an understanding of how widely applicable these you know these data are as of today.

(Refer Slide Time: 00:50)



Lecture 4

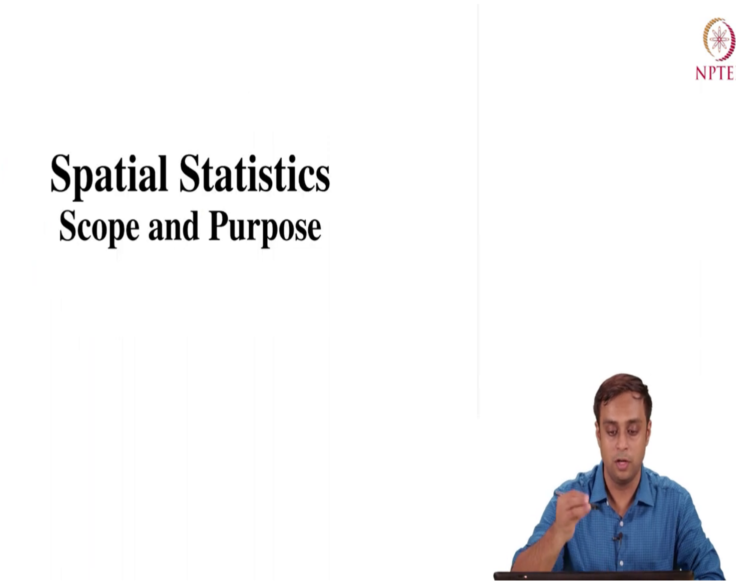
Recap

- Spatial Data and their applications
- Data Generating processes
- Spatial Data Structures
    - RASTER DATA
    - VECTOR DATA

We have looked at the data generating process which is specifically the remote sensing technology, the satellite sensor technology and so on and so forth. We also looked at the GIS which is a you know a computer software and hardware system to store, manipulate, analyze, visualize spatial data. Then in the last lecture we looked at what are called as the spatial data structures.

These are you know in particular the raster data structure and the vector data structure and we talked about how you know these are different and how their applications differ from problem to problem right.

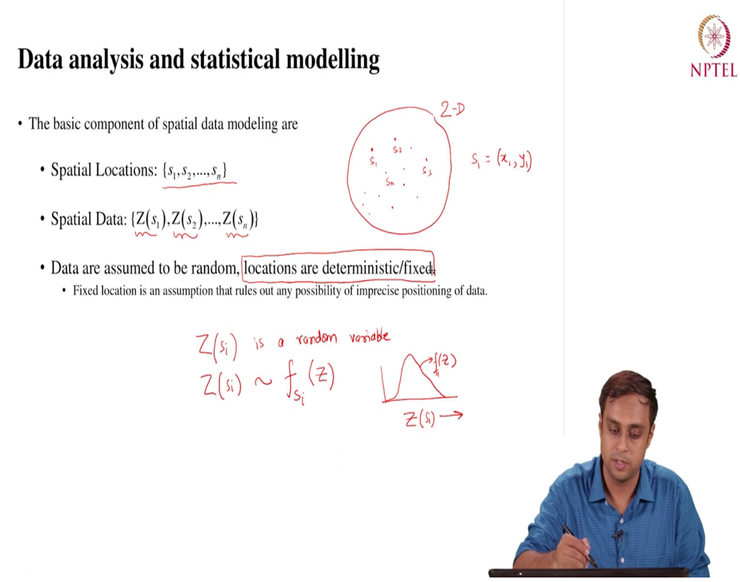
(Refer Slide Time: 01:33)



Spatial Statistics
Scope and Purpose

So, today we will move you know more formally to the statistics part and start departing from the data you know part of this course. So, we will today we will formally talk about spatial statistics its scope and purpose. So, let us get started.

(Refer Slide Time: 01:50)



Data analysis and statistical modelling

- The basic component of spatial data modeling are
  - Spatial Locations: $\{s_1, s_2, \ldots, s_n\}$
  - Spatial Data: $\{Z(s_1), Z(s_2), \ldots, Z(s_n)\}$
  - Data are assumed to be random, locations are deterministic/fixed.
    - Fixed location is an assumption that rules out any possibility of imprecise positioning of data.

$Z(s_i)$ is a random variable

$Z(s_i) \sim f_{s_i}(z)$

So, the first most primary basic component of spatial data modelling are; a, spatial locations. So, you know if you have any kind of spatial data you need to formally articulate what the locations are right. So, here what I have is a set of discrete locations s_1 to s_n right. So, data in a space in a given space or domain are located at locations $s_1, s_2, s_3, s_4, s_5, s_6$ and so on till sn right. Each location will have a coordinate system.

For example, if this is a two dimensional you know domain then you know s_1 will have a x_1 coordinate and y_1 coordinate also known as latitudes and longitudes you know in remote sensing science right. And then at these locations we actually observe spatial data. So, these spatial data are then articulated as value Z observed at location s_1, value Z observed at location s_2 and so on value Z observed at location s_n.

So, the uniqueness or the distinctness of spatial data that is observed at different locations is derived from the virtue of the location itself; that is the you know basic component of spatial data analysis right. And data are assumed to be random.

So, at each location what we assume is that you know you could observe a range of you know values of Z and you know these range can be characterized by a probability distribution function right. This is something we will cover in more detail in later today or in a in the next lecture.

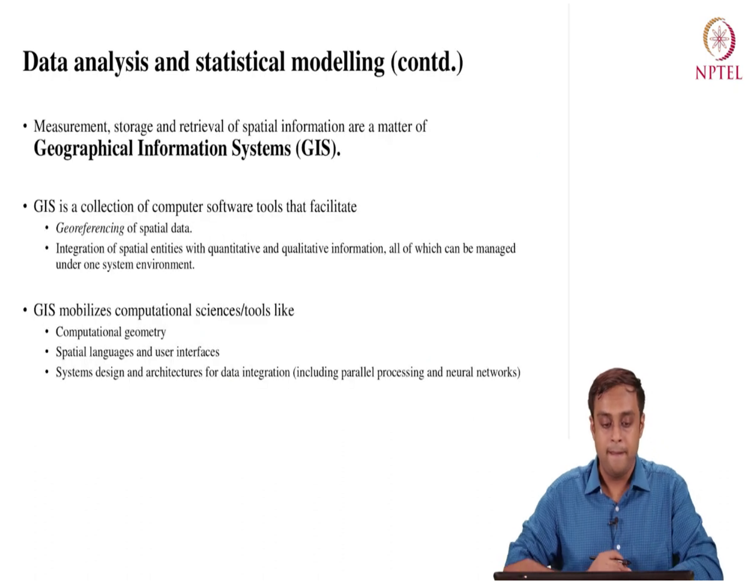
But the idea is that $Z(s_i)$ is itself a random variable right. And when we characterize a random variable we usually will you know we will use a probability distribution function which is to say there is a PDF f linked to location si which tells me what is the range of values that this random variable Z can take at location s_i right.

So, there will be a probability distribution at si which will determine the values that $Z(s_i)$ can take that we mark on the x axis again we will look at these things with much more detail going forward ok. And s_i is one of the locations you know that are there in my two-dimensional space that I am concerned about. Further and very importantly for this particular course we will take locations to be deterministic or fixed.

So, we do not take we do not consider any kind of uncertainty in the location itself. So, if you observe let us say for example, groundwater you know levels at location s_1 we say that you know ground water case groundwater levels by itself is a random variable, but the location has no uncertainty right it could not be it could not be s_1 plus minus delta right it is indeed exactly precisely at s_1 that we observe $Z(s_1)$.

Z by itself at s_1 is random. So, ground water level is random, but the location is fixed right. This is an assumption right this is an assumption you can have you can have uncertainty in the positioning of data as well, but that we deem out of scope for this particular course right. So, we will take locations at fixed and we will take data that is the observation the realizations at each location to be random variables.

(Refer Slide Time: 06:09)



Data analysis and statistical modelling (contd.)

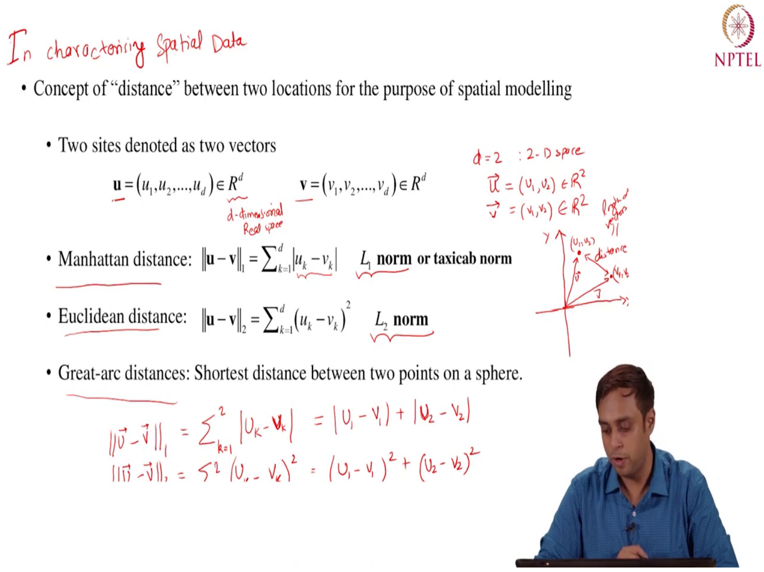
- Measurement, storage and retrieval of spatial information are a matter of **Geographical Information Systems (GIS).**
- GIS is a collection of computer software tools that facilitate
  - *Georeferencing* of spatial data.
  - Integration of spatial entities with quantitative and qualitative information, all of which can be managed under one system environment.
- GIS mobilizes computational sciences/tools like
  - Computational geometry
  - Spatial languages and user interfaces
  - Systems design and architectures for data integration (including parallel processing and neural networks)

So, this is something we have talked about at length. So, one component of you know spatial data is measurement storage retrieval through GIS system, you know. GIS is a collection of computer software tools that facilitate geo referencing of spatial data we understand these things now. It also sort of facilitates integration of spatial entities with qualitative and quantitative information and then all of that can be managed in one environment.

It mobilizes computational sciences, specifically computational geometry, spatial languages and user interfaces something that we will look at you know all of these components we will study. But this is about you know management storage and so on we are now going to be more concerned about the statistical characterization of these data right. So, let us move to that.

But when we talk about spatial data this is an important component we were spent quite a bit of time on this and now we are moving to the statistical understanding of these data.

(Refer Slide Time: 07:09)



In charactorising Spatial Data

- Concept of "distance" between two locations for the purpose of spatial modelling
  - Two sites denoted as two vectors

    $\underline{\mathbf{u}} = (u_1, u_2, \ldots, u_d) \in \underbrace{R^d}_{\text{d-dimensional Real space}}$ $\quad \underline{\mathbf{v}} = (v_1, v_2, \ldots, v_d) \in R^d$
  - Manhattan distance: $\|\mathbf{u}-\mathbf{v}\|_1 = \sum_{k=1}^{d} |u_k - v_k|$ L_1 **norm or taxicab norm**
  - Euclidean distance: $\|\mathbf{u}-\mathbf{v}\|_2 = \sum_{k=1}^{d} (u_k - v_k)^2$ L_2 **norm**
  - Great-arc distances: Shortest distance between two points on a sphere.

$d = 2$: 2-D space

$\vec{u} = (u_1, u_2) \in R^2$

$\vec{v} = (v_1, v_2) \in R^2$

$$\|\vec{u} - \vec{v}\|_1 = \sum_{k=1}^{2} |u_k - v_k| = |u_1 - v_1| + |u_2 - v_2|$$

$$\|\vec{u} - \vec{v}\|_2 = \sum^{2} (u_k - v_k)^2 = (u_1 - v_1)^2 + (u_2 - v_2)^2$$

So, in order to characterize, in characterizing spatial, you know data right. So, in characterizing spatial data, the concept of "distance" between two locations is fundamental right. So, we cannot really model spatial data unless we are able to measure distance between any two given pair of locations where we observe data right.

So, here what we do is we give you a most general form of two locations denoted as *u* and *v* right and these are vectors right. So, and these are vectors in d-dimensional space right. So, this is, these are vectors in d-dimensional real space right we can always whenever we look at this generalization to d-dimensional space we can always learn or figure it out further by putting $d = 2$ that is a two-dimensional space something we understand very well right.

So, how do I visualize *u* and *v* in a two-dimensional space. So, now, my vector *u* which is bold in the typed out text I am just using a top arrow to denote that it is a vector it will have two components u_1, u_2 which will be in R^2. So, u_1, u_2 are nothing but XY coordinates or latitudes and longitudes of the data and now location *v* is given by v_1 and v_2 again these are just XY coordinates at this second location which will lie on the second you know a two dimensional real space right.

And in this two dimensional real space where you have *x* and *y* right we can have vector *u* given as given by (u_1,u_2) from the origin this is my vector u and vector v could be given by this joining the origin to a point which is characterized by the coordinates (v_1,v_2). Ok. Given these the question is how do I figure out the distance between these two spatial locations right.

Unless we are able to measure this distance we will not be able to characterize what is the dependence how dependent are these two you know the observations that we see or the random variables that we see at these two locations are.

So, to be able to characterize that we need to characterize distance. The most popular three metrics for doing that is called as the Manhattan distance, the Euclidean distance and the Great Arc distance right. I am sure some of you must have heard of these earlier.

So, the Manhattan distance is nothing but a summation of the absolute difference of each coordinate values at these locations. So, for our example the Manhattan distance *(u - v)*, Manhattan distance will be given as:

$$Manhattan\,Distance\colon\ ||u - v||_1 = \sum_{k=1}^{2} \left|u_k - v_k\right| \qquad L_1\,Norm\ or\ Taxicab\ Norm$$

Right. Which is equal to nothing but absolute difference between u_1 and v_1 plus the absolute difference between u_2 and v_2 ok.

$$\left|u_1 - v_1\right| + \left|u_2 - v_2\right|$$

Similarly, the Euclidean distance which is also known as the shortest path right. That is given as the which is called as the L_2 norm can be written as following in the two dimensional space.

$$Euclidean\,Distance\colon\ ||u - v||_2 = \sum_{k=1}^{2} (u_k - v_k)^2 \qquad L_2\,Norm\ or\ Shortest\ Path$$

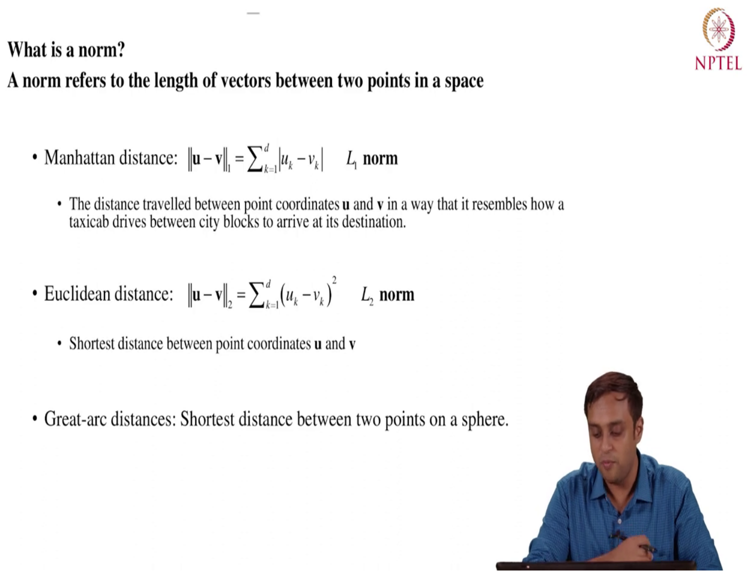
Which is nothing but

$$(u_1 - v_1)^2 + (u_2 - v_2)^2$$

Ok. Now let us move forward and visualize what these distances mean.

So, first of all I am giving them these characterization L_1 norm and L_2 norm and I want to sort of now formally define what a norm really is.

(Refer Slide Time: 11:47)



What is a norm?

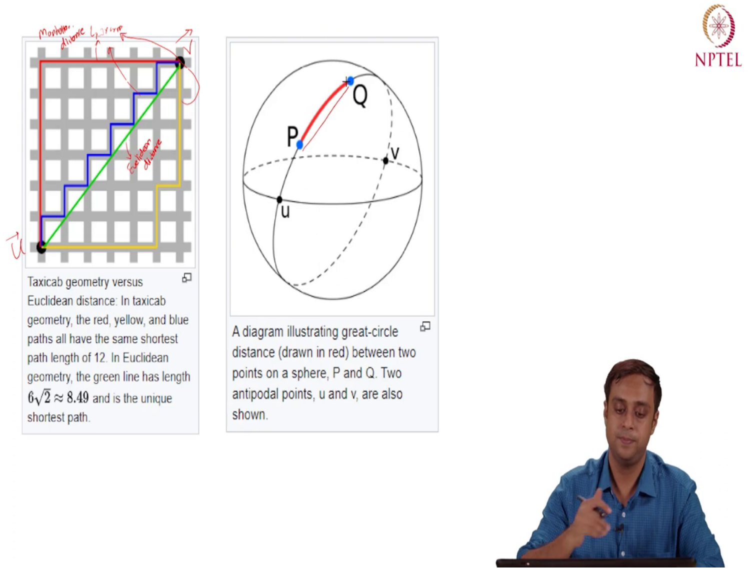
A norm refers to the length of vectors between two points in a space

- Manhattan distance: $\|\mathbf{u}-\mathbf{v}\|_1 = \sum_{k=1}^{d} |u_k - v_k|$ L_1 **norm**
  - The distance travelled between point coordinates **u** and **v** in a way that it resembles how a taxicab drives between city blocks to arrive at its destination.
- Euclidean distance: $\|\mathbf{u}-\mathbf{v}\|_2 = \sum_{k=1}^{d} (u_k - v_k)^2$ L_2 **norm**
  - Shortest distance between point coordinates **u** and **v**
- Great-arc distances: Shortest distance between two points on a sphere.

A norm reference refers to the length of vectors between two points in space. So, let us go back to our characterization. So, we had points u and v in front of your screens and the distance between them is the length of vector between these two points. So, the norm is a formal quantitative measure of the distance between these two points.

The Manhattan distance is the distance travelled between point coordinates u and v in a way that it resembles how a taxicab would drive between city blocks to arrive at its destination starting from point u to point v.

(Refer Slide Time: 12:29)



Taxicab geometry versus Euclidean distance: In taxicab geometry, the red, yellow, and blue paths all have the same shortest path length of 12. In Euclidean geometry, the green line has length $6\sqrt{2} \approx 8.49$ and is the unique shortest path.

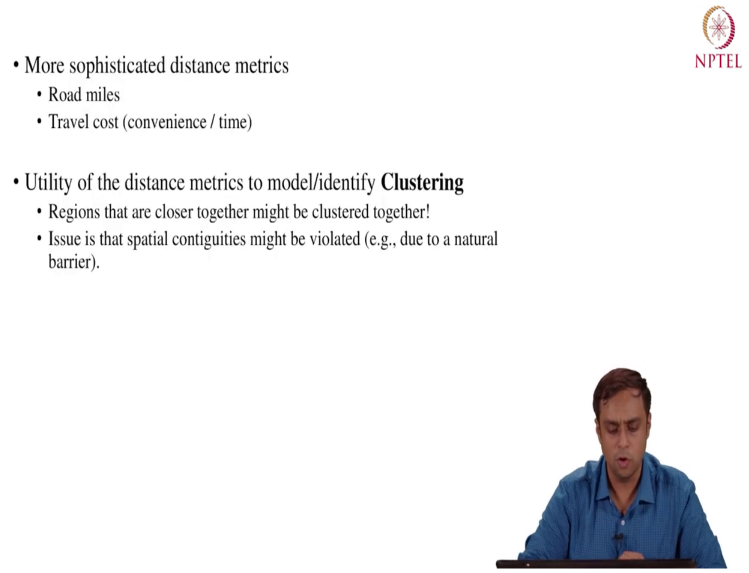
A diagram illustrating great-circle distance (drawn in red) between two points on a sphere, P and Q. Two antipodal points, u and v, are also shown.

So, let us go back and visualize this. So, on your screens on the left you have a picture which shows, you know, a red line, a blue line and a yellow line all of which are alternative routes that a taxicab would take between city blocks to arrive from point u to point v ok.

So, that is what Manhattan distance really is able to calculate. The second one the second type of measure for distance between any two given points in space is called as a Euclidean distance and this is indeed the shortest path between the points *u* and *v*. So, again go back to the picture the green line is what we call as the Euclidean distance right. So, this is the L_2 norm and the green blue and yellow lines as I said earlier are called L_1 norm or they are termed as the Manhattan distance, the Manhattan distance ok.

The third kind is called as the Great Arc distance which basically takes into account the fact that the shape of earth is a sphere. So, if I am moving from point *u* to *v*, I would have to move on a arc rather than a straight line between *u* and *v*. So, this distinction gives us yet another alternative measure of distance between two points in space which is called as the Great Arc distance ok.

(Refer Slide Time: 14:11)



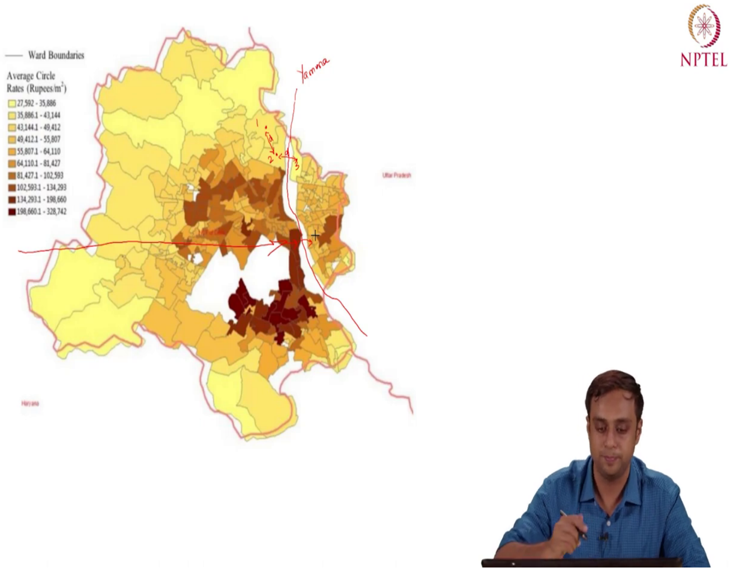
- More sophisticated distance metrics
  - Road miles
  - Travel cost (convenience / time)
- Utility of the distance metrics to model/identify **Clustering**
  - Regions that are closer together might be clustered together!
  - Issue is that spatial contiguities might be violated (e.g., due to a natural barrier).

So, these are primitive you know distance metrics. So, you know they we have we have more sophisticated you know distance metrics I mean you can imagine in India the taxicab characterization between city blocks would not really work right. So, you will have more sort of complicated sophisticated pathways that a taxicab would take if they are following what let us say a Google maps route right.

So, road miles that they cover or road kilometers that they would cover provides us yet another metric of distance between two points u and v in space. Another method is called as the travel cost method. So, now, travel cost you know accounts for not only distances, but also things like time and convenience. So, you can have issues like traffic congestion rights all those concepts can also be sort of you know brought into defining distance between two locations. So, that these are more sophisticated measures.

But the question is what is the utility of distance matrix right. So, the utility really is to be able to model or identify clustering between different pairs of points or locations in space where we are able to observe a data. Here usually regions that are closer together might be clustered together right and you know and that sort of gives us a point forward in terms of understanding or characterizing a spatial dependence in data.

(Refer Slide Time: 15:44)



Let us look at this example of the real estate market of Delhi. So, the map that you see is a map of average circle rates which is rupees and meter squares where you have different districts in the national capital territory of Delhi. The light yellow colors are referring to a low average circle rates in a given district that is land is relatively inexpensive in those areas. And the darker circles darker sort of you know blocks or polygons are the ones where the land is priced you know at an expensive rate right.

So, what you see here is that let us say if we consider district 1, district 2 and district 3 on this map. The distance between district 1 and 2 would be likely sort of to be will be calculated between the center of mass of these districts these districts respectively right.

So, the center of mass of district 1 versus center of mass of district 2 will have certain distance length it can be calculated using the Manhattan length the Euclidean length or Euclidean distance or the great arc distance or even more or more sophisticated measures right.

Similarly, two and three also will have a distance calculated between them you know between their center respective center of masses right. Now what we see here on this picture when we look at it seems like the distance between 1 and 2 is quite similar to distance between 2 and 3. However, 2 and 3 encounter a natural barrier in terms of the river Yamuna right.

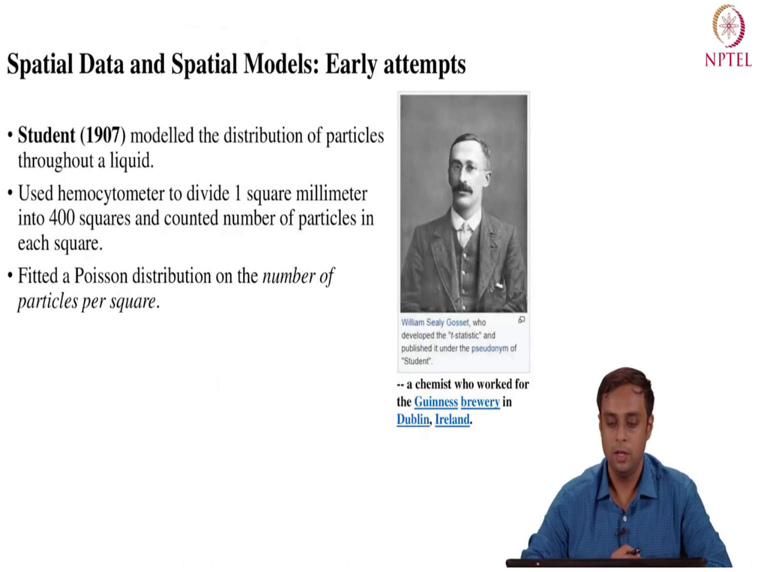
So, this natural barrier will require us to have a more sophisticated distance metric than a direct calculation of a Euclidean distance between points 2 and 3 right. So, this natural barrier what it can do is it can cause reversal of trends in terms of the real estate values. Something like that we see if we look at the map of Delhi we come from west to east when we come from west to east we see that the real estate prices are rising.

But as soon as we hit we hit the natural barrier and we go on the other side the prices are drastically lower right that sort of reversal can come from a natural barrier which is also a geographic entity which we want to sort of try and model you know through the tools that we learn in this course.

So, what I am trying to really say is that distance between two entities is a fundamental metric of how we can study you know a spatial data, dependence between them things like clustering and so on and so forth and how we measure distance can be highly sophisticated rather should be sophisticated enough.

So, that we can actually model the real world you know real world observations that we see an example of which here is that the natural barrier can actually cause a reversal of trends in terms of the real estate values in New Delhi. This is you know a real world data providing us a sort of a you know a motivation to think about distance very carefully when we work with these data ok.

(Refer Slide Time: 19:14)



Spatial Data and Spatial Models: Early attempts

- **Student (1907)** modelled the distribution of particles throughout a liquid.
- Used hemocytometer to divide 1 square millimeter into 400 squares and counted number of particles in each square.
- Fitted a Poisson distribution on the *number of particles per square.*

William Sealy Gosset, who developed the "t-statistic" and published it under the pseudonym of "Student".

-- a chemist who worked for the Guinness brewery in Dublin, Ireland.

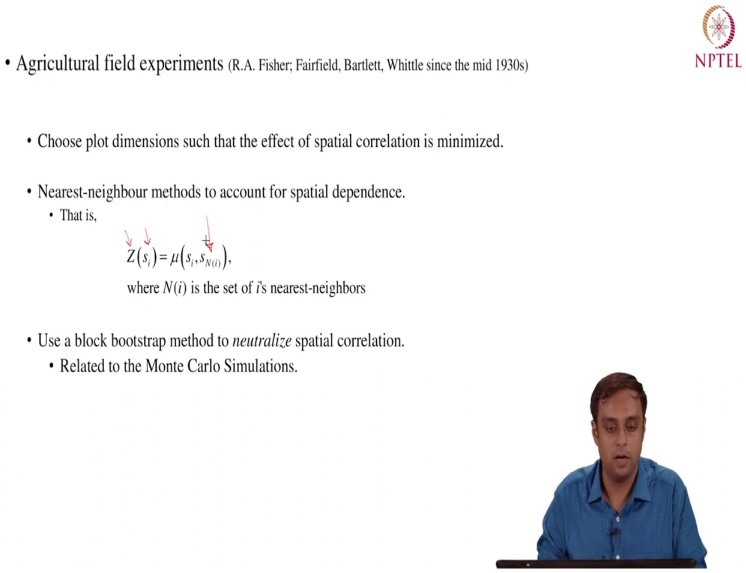
So, let us come to spatial data modelling. So, the earliest attempt one can you know that one can think of in terms of spatial models was done by a student in 1907 who modeled the distribution of particles throughout a liquid. He divided 1 square millimeter area or volume 1 square millimeter area sorry into 400 squares and counted number of particles in each square.

So, he basically took a flask of liquid he took a cross section of 1 square millimeter 1 millimeter squared, he divided it into grids 400 you know cells of equal sizes and started counting you know number of particles in each square. So, what you have is a count variable. So, you have a count of number of particles in square 1, in square 2, square 3, square 4 and so on in 400 squares.

So, when he modeled it right when he modeled it he had to be careful about where is the square that he is counting the number of particles in. Are those squares on the edges of this one square millimeter cube you know area domain that he is counting his you know particles in? Or is that you know in somewhere in the middle? Does he see higher density of count or higher counts in the middle or he sees higher counts in the on the edges? Right.

So, you can have higher clustering of points in the middle or higher clustering of points on the edges or vice versa right. So, all of a sudden you know this exercise that was done in 1907 has a documented you know characterization of space although not as sophisticated in terms of you know mathematical and statistical tools as we do today, but people were trying to already understand you know these methods at that time.

(Refer Slide Time: 21:20)



- Agricultural field experiments (R.A. Fisher; Fairfield, Bartlett, Whittle since the mid 1930s)
  - Choose plot dimensions such that the effect of spatial correlation is minimized.
  - Nearest-neighbour methods to account for spatial dependence.
    - That is,

$$Z(s_i) = \mu(s_i, s_{N(i)}),$$

where $N(i)$ is the set of i's nearest-neighbors

  - Use a block bootstrap method to *neutralize* spatial correlation.
    - Related to the Monte Carlo Simulations.

The other area where special spatial models have been there you know for a while now is agricultural field experiments. So, agricultural field experiments are done to estimate yields that is output per unit area on agricultural lands. And there you can imagine that you know you can have two farms together growing let us say wheat.

The kind of output per unit area that you are going to expect to observe on these farms is going to be quite similar right because they are side by side what would change they will experience similar, rainfall similar sunlight, they will have similar soils the farmers are likely to probably come from you know similar backgrounds I mean it is possible that the one farmer is richer than the other.

Let us assume that they are they have you know similar social economic you know background cultural influences and so on and so forth right. It is quite possible that plots that are located nearer in space will have similar yield values than plots that are located farther away right.

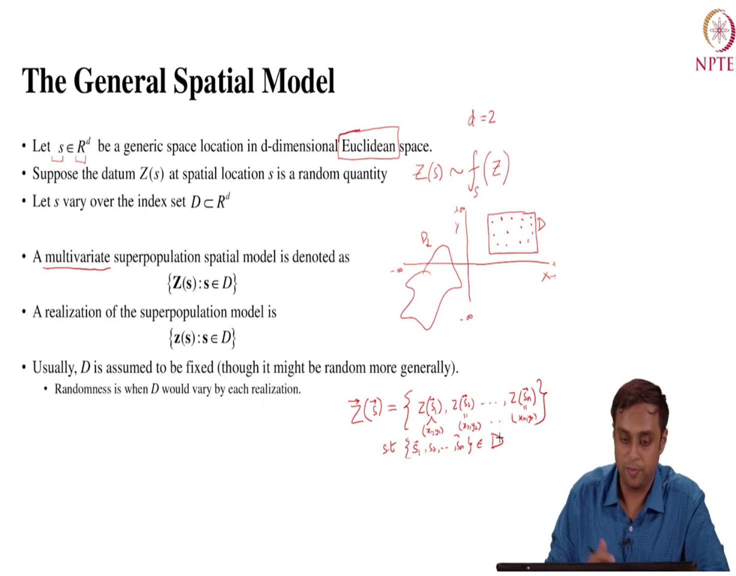
So, researchers have been actively choosing plot dimensions over, which they conduct these experiments strategically in order to account for a spatial dependence or spatial correlation in a way that they have distinct understanding when they move from plot 1 to plot 2 if these plots are large enough. It is likely that you are going to have a distinct understanding of yield levels when you move from plot 1 to plot 2.

Otherwise if you do not do that you can have 100 a 100 plots highly correlated with each other you are really not learning much about what is happening on average in a region right.

Then there are methods like nearest neighbour methods that accounts for spatial dependence we will talk about those in detail later I do not want to you know you can take a look at this. The only thing that I want to say is that the value that is observed at location i depends on what is happening at location N_i where N_i is simply a set of neighbours of i right.

There are other statistical techniques like block bootstrap methods to neutralize spatial correlation in data. These are things that you will learn later , but I just wanted to put it out there for your notes ok.

(Refer Slide Time: 23:43)



The General Spatial Model

- Let $s \in R^d$ be a generic space location in d-dimensional Euclidean space.
- Suppose the datum $Z(s)$ at spatial location s is a random quantity
- Let s vary over the index set $D \subset R^d$
- A multivariate superpopulation spatial model is denoted as
$$\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D\}$$
- A realization of the superpopulation model is
$$\{\mathbf{z}(\mathbf{s}) : \mathbf{s} \in D\}$$
- Usually, D is assumed to be fixed (though it might be random more generally).
  - Randomness is when D would vary by each realization.

NPTEL
d = 2
$Z(s) \sim f_s(z)$
D_1
D
y
x
$-\infty$
$\vec{Z}(\vec{s}) = \{ Z(\vec{s}_1), Z(\vec{s}_2) \dots, Z(\vec{s}_n) \}$
(x_1,y_1) (x_2,y_2) (x_n,y_n)
s.t $\{\vec{s}_1, s_2, \dots \vec{s}_n\} \in D$

So, let us come to a general spatial model. A general spatial model first of all has a space generic space location it in d-dimensional Euclidean space ok. So, when we say this we have said quite a bit first of all we have said that location will be indexed by s right. s will be in will be part of a d-dimensional real space whenever we see this generalized D what we have learnt till now we just set D equals 2 and we start to develop our concepts from there.

So, we can have a generic location parameter s which identifies different points in a two dimensional space with coordinates x and y in a in R^2 space right and what we are also saying is that we are going to work with a Euclidean space. So, now, when we specify or we declare a Euclidean space what we are also declaring is our method of measuring distance between two points right.

So, we have said quite a bit in that one sentence right now suppose the datum which is the data point Z in location or at location s is a random quantity. Now we have seen what we mean by a random quantity is that Z of s is going to be distributed by a probability distribution function which is specific to that location s and allows us to draw different realizations of Z right. And s can vary over the index set or domain D in the d-dimensional space.

So, again I will I am going to set D equals 2, I am simply working with a given specific domain d. So, of course, the real the real two-dimensional real space is infinitum of the x coordinate and the y coordinate jointly put together as orthogonal axes right. So, I have x which goes from minus infinity to plus infinity there is y axis it goes from minus infinity to plus infinity. I define a domain D which is my area of analysis alright.

So, I am going to define a domain D in this area of course, it does not have to be a rectangle it does not have to be a square it can be any irregular polygon that you can imagine. So, I can also sort of draw a more complicated domain D for your understanding and s is only allowed to vary in this D ok. So, at a time we are going to work with a given domain in this d dimensional real space to understand concepts again we are setting D equals 2.

Then a multivariate spatial model a multivariate spatial model. So, we have used the term multivariate. So, we have work going to work with multiple variables and the variables that we have are the random variables. So, multiple random variables are going to be denoted in a spatial model as following, you are going to have Z as a vector which is in bold. Given with the index location s which is itself also a vector because it can be a two-dimensional vector, a three-dimensional vector depending on s being in what type of a d dimensional space right.

And s must belong to D. So, we are going to restrict our analysis or our model of the data characterization of the data to D to this set D ok. So, what we are saying is that this Z capital sorry bold s bold basically which are vectors what we are saying is that these are a multivariate vector.

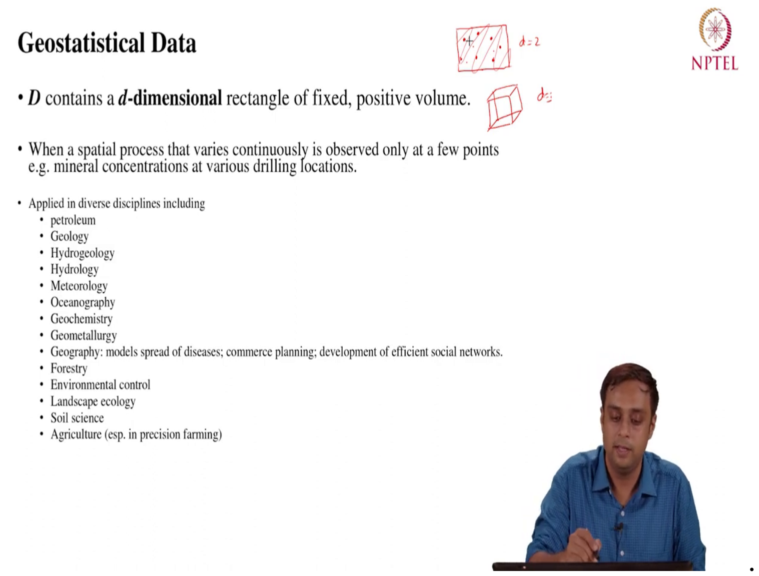
So, you have multiple variables. So, you have Z at location s_1 you have Z at location s_2 so on Z at location s_n ok. And each s itself is a vector which basically means that this s is x, y; x_1,y_1, s_2 is x_2,y_2 and keep going and s_n is x_n , y_n.

Such that s that is s_1, s_2, keep going s_n must lie in the domain of interest right. So, the analyst at least gets the choice to work with the domain that they want to work for right. So, you cannot really be working with for an indefinite space to begin with right. I mean you need to be able to fix a domain for which you are going to observe data right. You are going to you are not going to be able to observe data in a real space which is unbounded right.

That kind of analysis un is a un is tractable. So, we do not you know do that we fix d and we move forward from there. And so, these are all random variables and a realization in space is given by small z right from capital to small z and at every location s.

So, we only get to observe one entity one realization for the entire random variable we will formally look at what this really means, but a realization in space is given by small z right. And you know we have said that d is going to be a we are going to assume that d is fixed it could be random that, but that is out of scope for this particular course right ok.

(Refer Slide Time: 29:29)



Geostatistical Data

NPTEL

- ***D*** contains a ***d*-dimensional** rectangle of fixed, positive volume.
- When a spatial process that varies continuously is observed only at a few points e.g. mineral concentrations at various drilling locations.
- Applied in diverse disciplines including
  - petroleum
  - Geology
  - Hydrogeology
  - Hydrology
  - Meteorology
  - Oceanography
  - Geochemistry
  - Geometallurgy
  - Geography: models spread of diseases; commerce planning; development of efficient social networks.
  - Forestry
  - Environmental control
  - Landscape ecology
  - Soil science
  - Agriculture (esp. in precision farming)

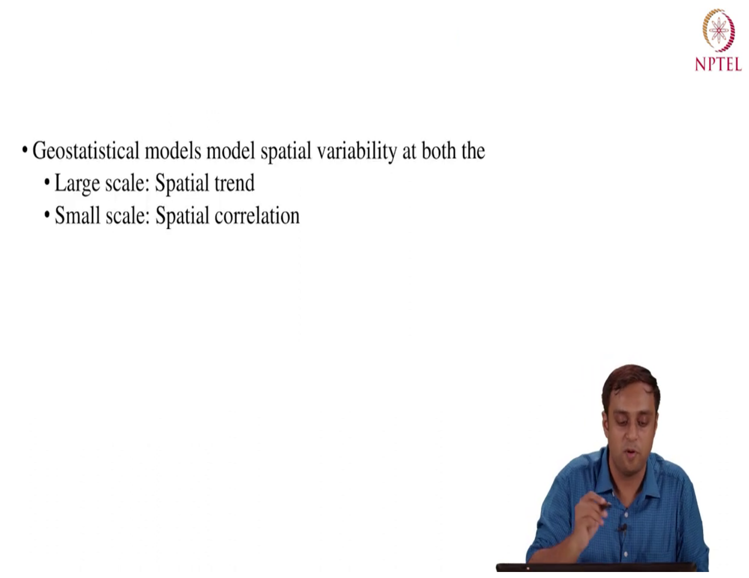
So, what are the different types of a data that would follow that given model? The most popular kind is called as the geostatistical data. Here D contains of a D contains a d dimensional rectangle of fixed positive volume. So, what we are looking at is a rectangle this would be a rectangle in d equals 2 and in d equals 3, it is going to be a it is going to be a three dimensional you know cube could be a d-dimensional or d-dimensional you know fixed positive volume domain that we are working with.

And when a spatial process that varies continuously is observed only at few points in space that is what is deemed to be a geostatistical data set. An example that I am giving here is mineral concentrations at different locations in space. So, the idea is that on land when I am trying to extract for example, I am trying to extract coal or on sea I am trying to sort of search for oil the point is that oil exists almost everywhere in this domain *D* right.

So, oil will exist at all locations in *D* right, but I can only monitor or observe the whether or not I can find oil or not or how much is the quantity what is the level or depth where the where I can extract oil from at only certain search locations right you know practically I can only probe certain locations in *D* right. Although I am probing only certain locations there are possible observations at every location that is possible in *D*.

So, you have infinitely many you know points in d that you can you can actually find the quantity or quality of oil that is available right. So, that is that is what is called as a geostatistical data set. I have given a wide range of applications that come with geostatistical data set. This is the most popular kind and we will most of this course in most of this course we will be looking at this type of a data set.

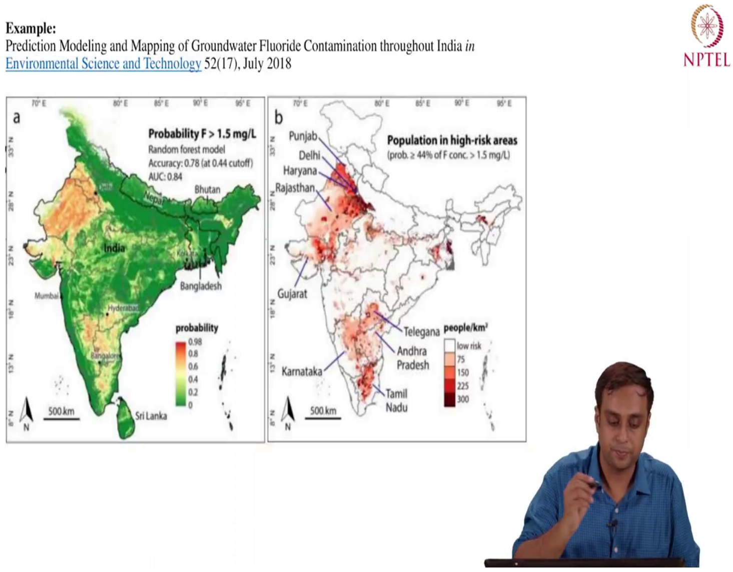
(Refer Slide Time: 31:52)



- Geostatistical models model spatial variability at both the
  - Large scale: Spatial trend
  - Small scale: Spatial correlation

So, again you know we have talked about spatial heterogeneity spatial dependence you know large scale spatial trends, small scale spatial correlation spatial dependence those types of understandings will start to directly apply to these data.

(Refer Slide Time: 32:08)



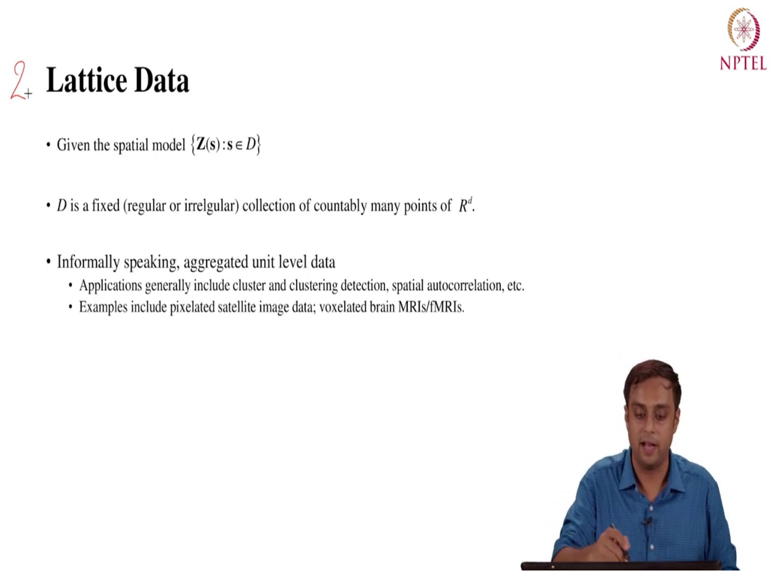
Example:
Prediction Modeling and Mapping of Groundwater Fluoride Contamination throughout India *in* Environmental Science and Technology 52(17), July 2018

The second kind so, there are here are some examples of geostatistical data one of the most prominent example is groundwater. Here is an example of groundwater quality that is fluoride contamination.

The idea is that look you will have ground water everywhere the question is where can you really you know observe it how wherever you dig a well you will be able to observe what is happening there in terms of quantity and quality, but that does not mean that wherever I have not dug there is no ground water there with ground water is available at continuum at the entire domain *D* of interest.

Here in this left picture the domain of interest is the entire you know conterminous region of India right as a country. On the left the region of interest sort of is let us say it is mostly you know states ranging from Punjab, Delhi, Haryana Rajasthan and Gujarat and then there are some states in south. So, you have a more restrictive region of interest that you are looking at on the right hand side. So, you can have both possibilities with these data.

(Refer Slide Time: 33:19)



2.

Lattice Data

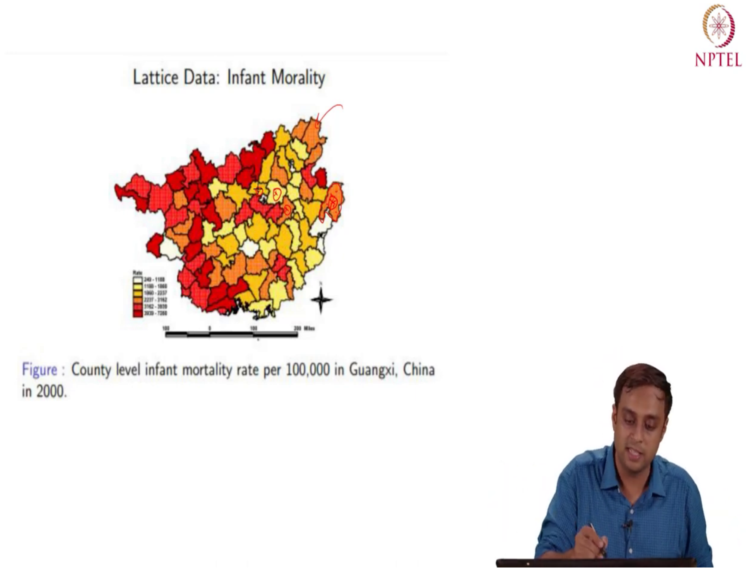
- Given the spatial model $\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D\}$
- D is a fixed (regular or irrelgular) collection of countably many points of R^d.
- Informally speaking, aggregated unit level data
  - Applications generally include cluster and clustering detection, spatial autocorrelation, etc.
  - Examples include pixelated satellite image data; voxelated brain MRIs/fMRIs.

The second type of you know spatial data are called as the lattice data. Again here D is a fixed regular or irregular collection of countably many points of the d dimensional real space. So, here I do not have entities moving continuously across throughout the space. I will have fixed given regular or irregularly spaced points where I can where I can observe data.

A very good example of this is states in India right. So, if I am observing GDP levels at state you know at for different states in India. I can only move through the centers of mass of you know different states in India right. I cannot really tell you what is happening within each state if I only have observation at the level of state GDP for India in 2021 right.

So, there if you have such administrative boundaries, country boundaries, district boundaries, taluk boundaries and so on and so forth. Some kind of an aggregated unit level data that is known as the lattice data.

(Refer Slide Time: 34:32)



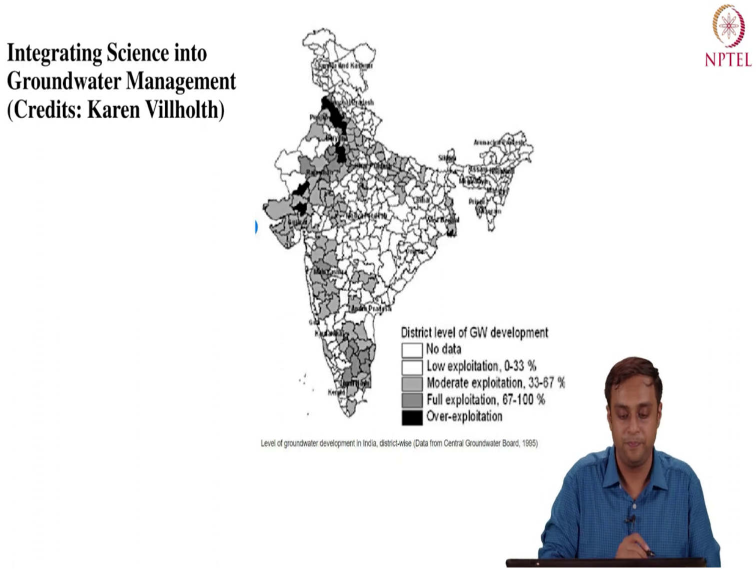
Lattice Data: Infant Morality

Figure : County level infant mortality rate per 100,000 in Guangxi, China in 2000.

And a very natural example is in front of you is a is infant mortality rate for a region in China in 2000 what you see here are irregularly spaced, irregularly shaped, but fixed and given administrative units that are districts in this region where you can observe infant mortality rate right. Given a within a given fixed you know region of observation you do not observe different levels of infant mortality. You have one level which is given at its center of mass.

So, you basically have discrete centers of mass where you can observe data right. So, that is called as this type of data set is called as the lattice data right.

(Refer Slide Time: 35:19)



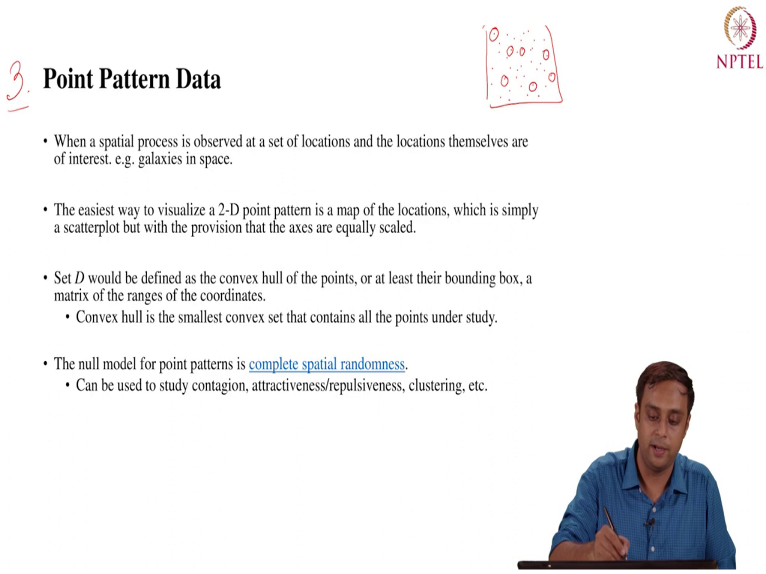
Integrating Science into Groundwater Management (Credits: Karen Villholth)

District level of GW development
No data
Low exploitation, 0-33 %
Moderate exploitation, 33-67 %
Full exploitation, 67-100 %
Over-exploitation
Andhra Pradesh
Kerala
Goa
NPTEL

Level of groundwater development in India, district-wise (Data from Central Groundwater Board, 1995)

So, this is something we have talked about now we have integrating groundwater management at district level right. So, there are many applications such as such the real estate data that we saw for the national capital territory of Delhi was a lattice data right. So, we have seen this and we have used this example multiple times throughout this course.

(Refer Slide Time: 35:38)



3. **Point Pattern Data**

- When a spatial process is observed at a set of locations and the locations themselves are of interest. e.g. galaxies in space.
- The easiest way to visualize a 2-D point pattern is a map of the locations, which is simply a scatterplot but with the provision that the axes are equally scaled.
- Set *D* would be defined as the convex hull of the points, or at least their bounding box, a matrix of the ranges of the coordinates.
  - Convex hull is the smallest convex set that contains all the points under study.
- The null model for point patterns is complete spatial randomness.
  - Can be used to study contagion, attractiveness/repulsiveness, clustering, etc.

The third kind of data is called as the point pattern data. When a spatial process is observed at a set of locations and the locations themselves are of interest. The you know are of interest in the sense are a random variable of interest.

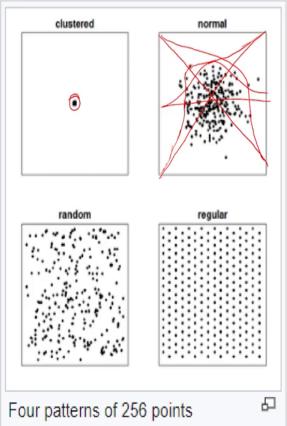
So, if the location themselves are a random variables of interest then those spatial processes are modeled as point patterns. The easiest way to visualize a two dimensional point pattern is a map of locations which is simply a scatter plot, but with the provision that axes are equally spaced right. Here thus domain *D* the capital *D* domain which with which we are working is basically a convex hull of points which are random variables of interest in space right.

And this convex hull is nothing but the smallest convex set that contains all points in our study right. So, this is the lot of technicality, but what I really want to emphasize here is that if the location itself is a random variable. For example, if I am studying deforestation and if I want to sort of if I have a wilderness of tree canopies and you have trees being cut at different locations in space right.

So, let us say if we have tree canopies at different locations you know in space which is given already and you want to study t forestation and the random variable of interest is for each point that we observe a tree you know a structure what is the time to deforestation right. So, what is the time at which you will observe deforestation. Then you know as soon as through time as soon as you observe a given tree being cut down you can mark it with a circle and that becomes a random variable of interest.

So, as trees are cut in space right you start observing a random pattern of how these point patterns are emerging or evolving through time. This is the specific kind of data is called as point pattern data.

(Refer Slide Time: 38:04)



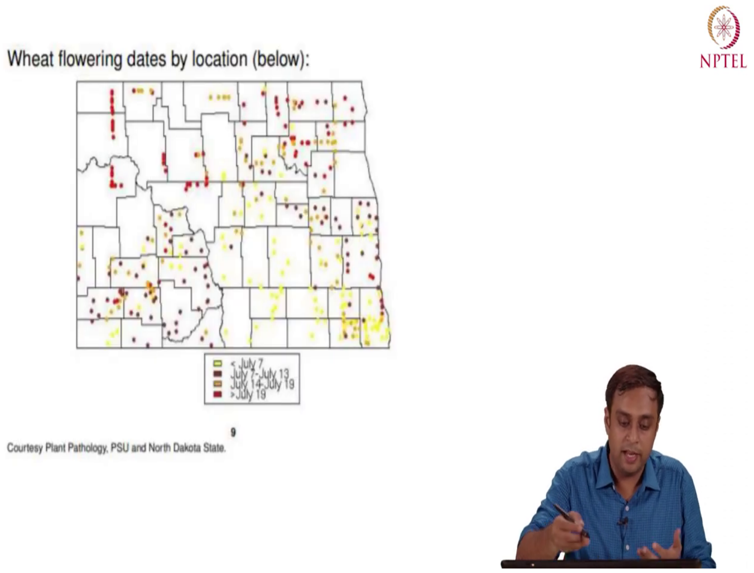
Four patterns of 256 points

Let me give you a some examples. So, you know you have you know points scattered in space which you can model as different distributions right. So, first is clustering right. So, you see the first you know box there is a highly clustered spatial data set you have a point pattern where all the points are clustered at one location right. So, points come together in a way that they are clustered in one location.

The second is normal. So, what it means is the points have come together in space in a way that they resemble normal distribution no matter what axis they are looking at. So, if I go by this diagonal in front of your screen you see a very you know you see no data you see no data at both extreme ends of this domain and as we move forward you start to have a higher count of points in space.

So, what this when we start to count them and plot frequency of counts what we see is a normal distribution along these two directions right. So, you can go in any direction and you will have this a pattern. Similarly you have a random pattern where you cannot really specify a distribution and then you have a regular pattern you know which is equally spaced points ok.

(Refer Slide Time: 39:32)



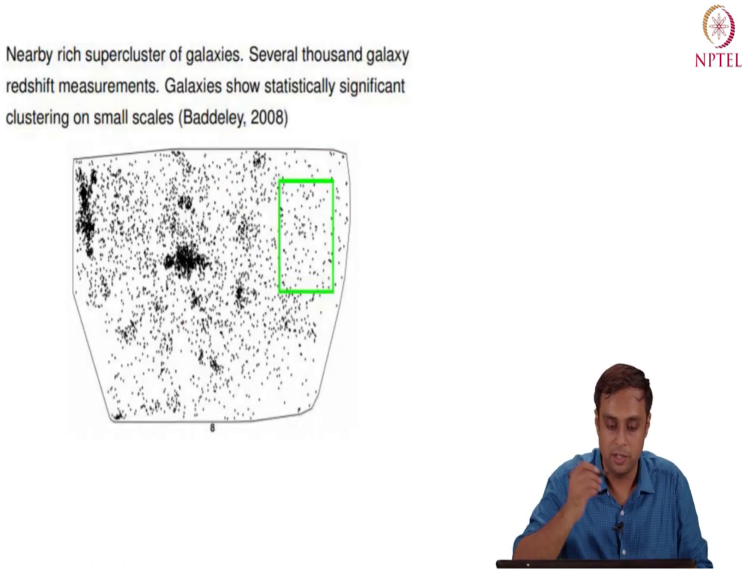
Wheat flowering dates by location (below):

9

Courtesy Plant Pathology, PSU and North Dakota State.

So, there are other you know other examples wheat flowering dates by location.

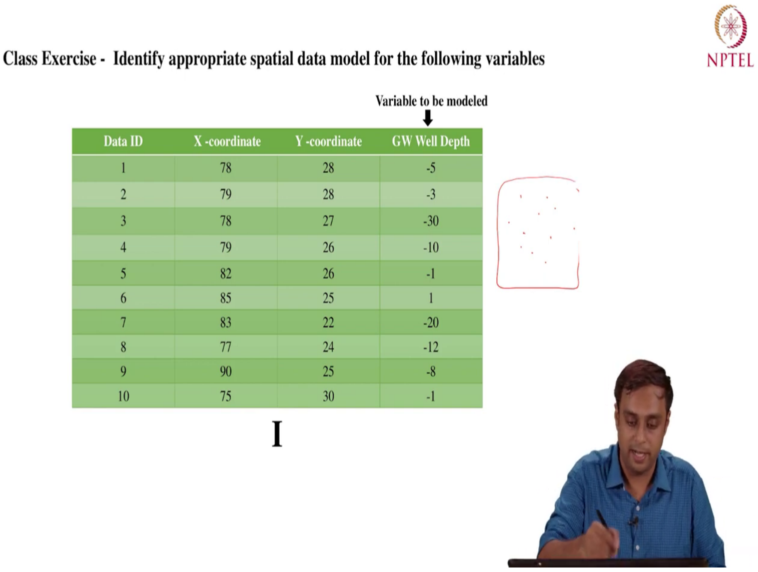
(Refer Slide Time: 39:38)



Nearby rich supercluster of galaxies. Several thousand galaxy redshift measurements. Galaxies show statistically significant clustering on small scales (Baddeley, 2008)

Then you have you know super cluster of galaxies you know how they arrange over space. So, you know physical scientists are very interested they use a spatial statistical tools to model locations or distribution of galaxies.

(Refer Slide Time: 39:55)



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

Variable to be modeled

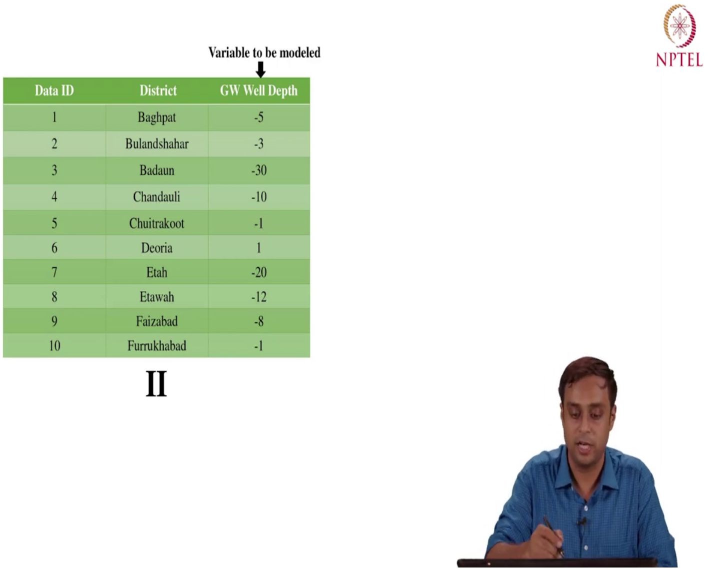
Data ID	X -coordinate	Y -coordinate	GW Well Depth
1	78	28	-5
2	79	28	-3
3	78	27	-30
4	79	26	-10
5	82	26	-1
6	85	25	1
7	83	22	-20
8	77	24	-12
9	90	25	-8
10	75	30	-1

I

So, as the last item in this class I want to sort of I want to sort of you know want you to identify appropriate spatial data model for the following variables of interest. So, in the first I have a ground water level depth given by given at 10 different locations all at all these locations I have a X coordinate and a Y coordinate.

So, at all given locations I have a X coordinate and a Y coordinate I have 10 locations in all. So, I will have some kind of a domain I want to ask you what kind of spatial data model will you be able to put on it. Will it be a geostatistical data set? Will it be a lattice data set? Will it be a point patterns data set?

(Refer Slide Time: 40:40)



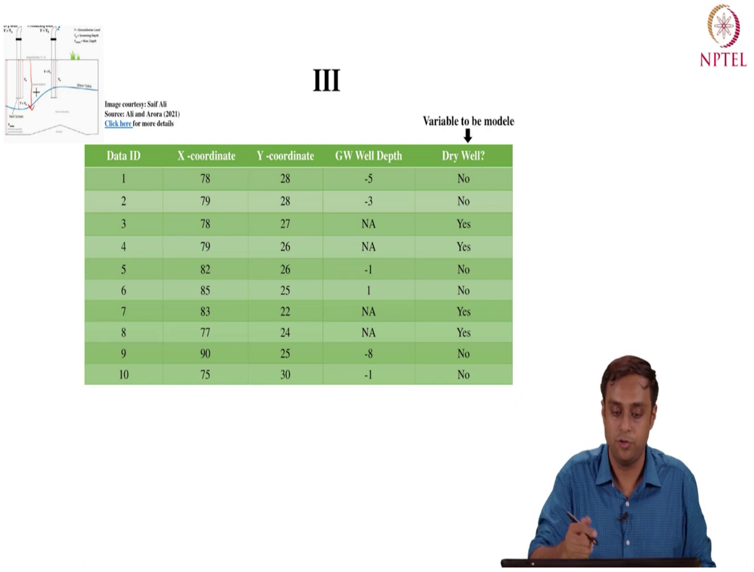
Variable to be modeled

Data ID	District	GW Well Depth
1	Baghpat	-5
2	Bulandshahar	-3
3	Badaun	-30
4	Chandauli	-10
5	Chuitrakoot	-1
6	Deoria	1
7	Etah	-20
8	Etawah	-12
9	Faizabad	-8
10	Furrukhabad	-1

II

The second example is again by district. Now we have ground water depth by different districts in Uttar Pradesh. So, these are districts in Uttar Pradesh. You want to I want you to tell me or to think about what you know what type of a geostatistical data set are we looking at.

(Refer Slide Time: 41:00)



III

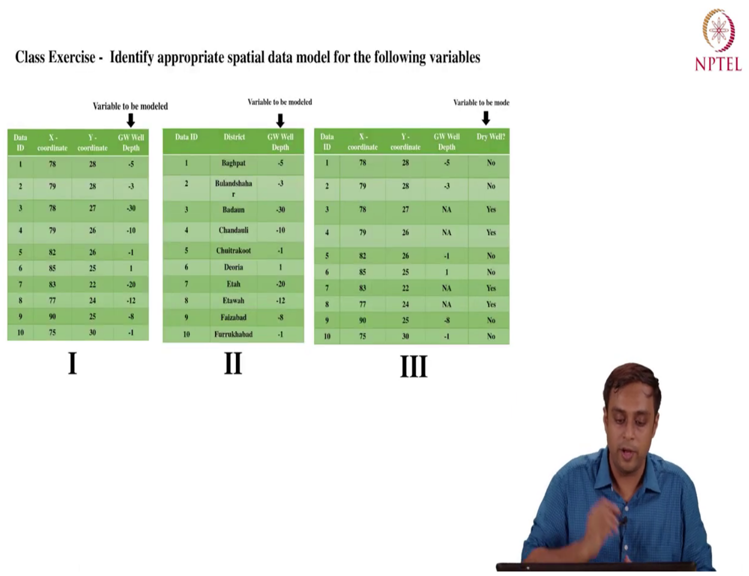
Image courtesy: Saif Ali
Source: Ali and Arora (2021)
Click here for more details

Variable to be modele

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

The third one we want to model whether or not we observe a dry well. So, what happens is that wells can go dry if ground water levels go deep enough. And we can only the model the variable of interest here is whether we have a dry well or we do not have a dry well right. If we want to model this what kind of a process are you going to use.

(Refer Slide Time: 41:24)



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

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

Variable to be modeled

Data ID	District	GW Well Depth
1	Baghpat	-5
2	Bulandshahar	-3
3	Badaun	-30
4	Chandauli	-10
5	Chuitrakoot	-1
6	Deoria	1
7	Etah	-20
8	Etawah	-12
9	Faizabad	-8
10	Furrukhabad	-1

II

Variable to be mode

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, in summary we have three different types of you know datas variables to be modeled. We have different types of data structures in an excel sheet and I want you to identify which type of spatial data model which will be the most appropriate for each of these. So, we will give you know some time to do that and then we will come back and we will resolve this query.