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

Lecture - 10A
Exploratory Spatial Data Analysis

Hi everyone welcome to the 10th lecture on Spatial Statistics and Spatial Econometrics. In today's lecture, we are going to sort of shift gears to working with data. And any time that you work with data the first step is exploratory analysis. So, what you really ought to do is to be able to summarize these data or explore you know conduct an exploratory analysis intelligently, so that you can start to discover patterns that are in these data right, in the data that you are working with. What do we typically do when we sort of do exploratory data analysis?

We begin by plotting histograms you know we try and look at the shape of the distribution, you know are the data symmetric in the distribution? are they skewed towards the right or skewed towards the left? That is to say are there more extreme outlier values on the right-hand side of the distribution or the left-hand side of the distribution right?

We also summarize by documenting the sample mean, the sample variance, the minima, the maxima, the 5th percentile, and the 95th percentiles. Things that you have heard of till now in this course you know already ok. We want to now sort of understand when we get data, that are geo-referenced that is to say that they have location coordinates. I mean they could be you know lat long coordinates, or they could be simply x y coordinates you know you may not be working with data that are necessarily geotagged. But they have some tag of a location, right?

Whenever you get data that are locationally or spatially delineated with whatever you know you know the spatial system that is used to create these coordinates for these data in the dimension that you are working in.

How do we go about conducting what is called exploratory spatial data analysis? Right. How would it be different from you know a typical exploratory data analysis? You know as a first step you know location is going to become you know a marker to which you are going to summarize the data right. So, now your means and your variances are not just sample means,

but they are sample means in the east-west direction versus the north-south direction and so on and so forth right.

Because we are studying data analysis we have to start with an example. So, let us do that and go step by step and look at what we look at when we try and sort of summarize you know spatial data, in terms of you know exploration.

So, here I want to sort of first of all you know, the example of the case that I am going to use to introduce exploratory spatial data analysis. Is adapted from the book written by Noel A Cressie which is a 1993 publication, it is also a reference book for this course. So, you know you can go back and you can read chapter two of this book you must read you know the first and the second chapter at least of this book. It is a very well-written book, however, it is a bit technical.

So, the job one of the jobs that I do through this course is to make this book a little bit more accessible to you as and when you start to read it right. So, let us look at the example that Cressie presents to explain to us what is exploratory data spatial data analysis and how to go about it. So, Cressie takes his example of coal ash data. So, what we are looking at is locations where we have samples of coal ash volume.

So, coal ash is the type of impurity when one wants to mine coal from beneath the ground, you do not really know how much coal can be found just right beneath where I am standing. And you know in this area you know what is the quantity of coal? what is the quality of coal? And one has to be able to estimate the quantity and quality characteristics of coal beneath the ground before one invests in you know pulling the earth out and engaging in mining. Remember, coal mining projects are you know often multi-decade projects, right?

So, when you go in you really are going to be spending a lot of capital in terms of you know mining that coal. So, what one does is that you know one has to explore whether or not there is coal beneath any particular region that we can look at. And what is the quality of that coal right, I mean you may have a lot of coal, but if it is you know not, so good quality. Perhaps you want to sort of prioritize mining in a different area where you have very high-quality coal right?

So, here is an example of a region that has been sampled for coal, right? And these crosses that you see on your screen are locations where you know coals are dug into the ground and

you know the earth is brought out so that you can get samples of coal that will provide you an idea of how much coal is available and of what quality right.

Now, the first thing that you notice here the first thing that we notice here is that we cannot get a sample of every location in space. That is to say that we cannot be digging holes at every location in space right, we can only dig these holes at locations where we see these crosses right? So, we only know what is happening beneath the ground at locations where we see these crosses. Everywhere else we do not see these crosses, you know where we have not to code or we have not sampled we are still blind to it right?

So, although I might get a very good idea of what is happening at this cross which I have circled with a red circle or some of these crosses. I will have no idea what is happening in the region around this these red-circled crosses. And this is one of the roles of spatial statistics this is one of the places where spatial statistics came into being is to be able to predict the quality and the quantity of coal beneath the ground where it has not been sampled. Using information from where it has been sampled that is the crosses by themselves.

So, the idea is that you know we cannot go out and sample every location in space, it is extremely costly, and it is almost nonsensical. It is one of the reasons why you know when India does its decadal census, you know it can only sample a certain percentage of the population it is not possible to go out and collect data or survey data from every individual in India or every household in India right? It is extremely costly and it's going to take a lot of time probably you will never be able to finish such an exercise.

So, we try and get a sample of some you know people and our sampling strategy is usually smart, and intelligent, it is driven by science, and using that sample we try and make inferences about the population ok. Similarly, in space here in this example, you are sampling certain points on the ground and you are leaving others out. But when you sample these points on the ground the idea, the purpose is that I should be able to predict what is happening in the region in the larger region using these many samples you know sample points on the ground right?

If nothing, I should at least be able to say what is happening in you know spaces within a smaller domain where I was not able to sample although I have a very good sample I think these are 294 sample points if you read the book. So, around 300 sample points, but that is

not sufficient I mean if I were to sample every point I mean there will be a very large value very very large value right?

You can dig here, you can dig here, you can dig here, you can keep digging you know as much as you would like alright? So that is where spatial statistics come in and then as a first step we want to sort of look at the data, summarize the data with the final objective that we will be able to predict what is happening in the region beneath the surface so that we can make these decision production decisions whether to go in or maybe move to a next location ok.

So, the coal mining industry oil exploration industry you know, and such you know is where you know most of the genesis of geostatistics really comes from ok.

So, here is another you know visualization of the data. This visualization is now happening in 3D. Now your west direction which we saw on the map is let us say it is the X axis and the north direction is let us say the Y axis. And the Z axis or the Z axis which is the height is giving me you know the content of coal ash that is found in the sample that was drawn at each location on the ground right.

So, my west is my X axis, north is my Y axis and I am looking at a height from the ground to sort of you know identify the percentage of coal ash. Of course, the larger the value of coal ash higher the impurity, and perhaps the lesser the desirable lesser desirable this you know coal mining project will become right.

So, let us move forward in the quest to summarize this data. So, in the next step what is being done is that this looks this orientation in which the data were originally you know procured, it is then sort of you know reoriented in our typical row versus column or x y you know the direction that we are very used to right. When we study you know the graphs we are very used to the X-axis and the Y-axis.

So, we sort of reorient the original data set to be represented with rows along the Y-axis. So, rows along the Y axis and columns along the X axis. Indeed the data were sampled in rows and columns right, they were just you know in an orientation in space that we could simply sort of you know to change a little bit and it gives us a very nice XY orientation to start working with.

Another thing in this representation is that instead of simply putting a cross what we are putting are, the values that represent the percentage of coal ash at that location right. So, in column 14 and row 23, the coal that was dug out or that was sampled had 9.99 percent of coal ash, right? Yet in column 1 row 14, you had a slightly higher percentage of coal ash which is 10.21 percent.

There are some locations where you know the coal ash percentage is quite high I mean we can go back and check I mean we saw a very high number here, which was closer to 17.6. Of course, that is to be found in the data set as well you know in column 5 and row 6 you have this value of 17.61 which is quite a high level of percentage coal ash. Especially, if you look at the samples around this particular location right. If you go one step north of this location you have 10.96, east is 10.87, south is 10.8 and west is 10.82. So, somehow 17.62 seems like an aberration, right?

So, straight away this inquisitiveness arises when we start to look at the data closely right. And we would want to then discover patterns in these data that you know that provide us you know information about the coal you know, that is available beneath the ground in the region of study right? Another thing one can do is that one could have a; one could have a sample mean. So, if I were given non-spatial data I will simply take all these values to sum them up and divided them by the total number of observations.

I will get; I will get a sample mean here you can get a mean which is row by row you can get a mean for row 23, for row 21, for row 20 all the way till row 1. And you can get a separate mean for column 1, column 2, column 3, column 4, column 5, column 6 and so on and so forth. So, now you are getting these mean values in the east-west direction and north-south direction which are different right? So, this is this right away introduces the specialty of the data, which is the spatial dimension of the data to summarize statistics.

Just like we are evaluating means for different rows and columns we might as well evaluate the mode, the median, the variance, the min, the max, and the 95^{th} and the 5^{th} percentile or other percentiles of our interest right? So, interestingly when we look at the data right away we can sort of you know see these dimensions play out.

So, as a next step you know what we start with is what we call detecting outliers and I am saying detecting outliers one, that is to say, that is the first step towards detecting outliers. What I have here is called the stem and leaf plot which perhaps was you know the traditional

version of a histogram. So, if you look at the stem and leaf plot the way it sort of turns out is skewed you know a bell curve, pardon me for my you know bad drawing.

But if you were to sort of take this bell this vertical stem and leaf plot and just rotate it to you know horizontal stem and leaf plot it will start looking like a density function or even something like a histogram correct? So, the stem and leaf plot first of all is just like a histogram the second is second thing is how is it actually constructed. So, the stem and leaf plot is constructed by taking each value in my data and rounding it off to the first decimal order. So, let us say I have 11.17 which will become 11.2, I have 9.92 which will just be 10, 10.21 which will be 10.2, and I have 10.14.

I am going to call it 10.1, then I have 10.82 which will be 10.8, 10.7, 9.9, 11.3 and so on and so forth. So, we collect all these values and we go from the smallest value say 6 and we go and figure out do I have 6 with decimal entities 6.1, 6.2, 6.3, 6.4, 6.5, and 6 point so on and so forth. I have in this entire data set you can search and you will find there are no values of percentage coal ash being as low as 6 percent right? So, the minimum that we see is 7 percent in this data.

So, in front of 6 we have nothing no entry in front of 7 when we move on to 7 we see 7 003. That means, in my data set I am going to have an entry that is going to be 7.0 another entry with 7.0, and then another entry with 7.3. I will have entries of three entries of value 7.6, one entry of 7.7, and then four entries or five entries are 7.8, and then two entries are 7.9 and you keep on going.

When you collect or organize or summarize your data in this fashion you have you know a stem which is all the values in front of the decimal point. And on the left-hand side of this, you know of this stem and leaf plot which is called the stem. And the values on the right-hand side which is the collection which is the frequency count using the decimal values the first-order decimal value on the right-hand side of the decimal point is called the leaf portion of the stem and leaf plot.

Like I said we can simply draw a histogram and we will get the same understanding of this entity the histogram that we build should be sort of binned with you know bin sizes of 0.5. So, I should have 6.0, 6.5, 7.0, 7.5, 8.0, 8.5 keep going all the way to 17.5 and 18 right? And now if you do want to do a frequency count and draw a histogram of these data on this value

you will get an exact representation that the leaf and; the stem and leaf plot is providing in the vertical orientation right?

So, you can replicate this it will also be helpful to understand what a stem and leaf plot is. So, using the stem and leaf plot what we can understand is that the density of these values density of these data seems to be skewed towards the right. So, you know you have a very high value of 17.6 then you have some values at 12 and 13.

And then most of the values are really sort of starting to pick up on the left-hand side at 11 you have a mean around you know 9 and 10. So, you have a large sort of distribution mass around there and then it starts to decline and this sort of dies out to the right ok. I have done I have not done a good job of plotting the histogram, but the density will be skewed to the right. I think I should redraw this sorry about that fumble.

So, the histogram if I were to draw it will look something like the following ok. So, you will have your 17 points you know 5 somewhere here and you will have your you know 6 here, your 7s here and so on and so forth ok. So, the mode and the median probably will be you know this one of the centrality measures will be around 9 or 10. It will probably be 9, I think that is where the height becomes the largest right? 9 or 9.5 will probably be where you will find the median of these data you know.

So, we know for a fact that the data are right skewed they are asymmetric of course. They are right-skewed, so they are asymmetric in nature. And you know the mean of coal ash value will likely be greater than the median because it is right skewed. So, a mean might sort of come out to be around I do not know 12 or 13 you know that is going to be mean, and the median is going to be shorter because it's right skewed you know density function. So, that is the first step.

Now, here clearly the values were so high, as high as 17.61 we located it seems to be a problematic value. So, I will mark it using the stem and leaf plot or a histogram I can mark the first potential outlier in my data ok.

So, I have marked the first potential outlier in my data, what is the next step, right? So, here I am just showing you in green circles you have all the values which I thought when I looked at this histogram or the stem and leaf plot for the first time, which were potential outliers things that did not agree with the larger region the areas that did not agree.

Now, these outliers can come up due to different regional reasons right I mean they could be because of measurement errors, you know they could indeed be outliers by themselves. But these are troublesome values that should raise a red flag when we calculate the mean value right. So, if we omit all these outliers and we calculate a mean it might be much lower than if you were to include all of them because they are all skewed in the right-hand side direction or the larger percentage coal ash direction.

So, what it really means is that if you calculate a sample mean and if you believe it blindly you might underestimate the coal quality found in that region. Because the percentage of coal ash or impurity will turn out to be much higher than probably, what it might seem if you were to remove these 5 6 outlier values. Especially the large one which is 17.61 ok.

Alright, so deducting outlier 2 is analyzing the difference between the mean and the median ok. So, when we looked at the histogram we just saw that if you have a skewed distribution the skewness comes from some values, which are sort of outside of the domain of the larger you know a large amount of data. Those are indeed troublesome values or outliers, right? and what is the signal of a skew in the distribution? It is the fact that the mean and median are no longer the same right.

So, if you have a symmetric distribution like in the bell curve for a normal distribution. Your mean and median are likely to be are going to be exactly the same for symmetric distribution mean is equal to the median right? So, that is the symmetric distribution, but in case you have an asymmetric distribution you have something like you know in red or let us say you have something like in blue, then you know you are no longer going to have your mean and median values to be higher right.

For blue, you are going to have your median be smaller than the mean and for red, you are going to have the mean be smaller than the median. So, comparing means and medians can help us detect outliers. You know it will be a signal again that maybe there are some outliers in the data that we should start to worry about ok. So, first of all, you know such. So, how does it sort of you know help us move beyond the histogram? I mean the histogram could also tell me that there is a skew right?

But you know moving to the second step serves a couple of purposes the first is that the median is a resistant statistic. So, measuring trends in data using medians is robust to any atypical observations or outliers right? So, first of all, the second step tells me that maybe you

should always you know summarize the median of a data set, whenever you summarize the mean of it. Because the mean might be you know misrepresenting or confounding due to some values which are very very high right.

For example, you know income levels in India if you look at the mean income level which is the average GDP or GDP per capita that is the mean GDP level. It might be much much higher than the median GDP level, because you know there are some you know individuals who may have a lot more, not a lot more wealth than you know all the other people in the country right?

So, sometimes you know mean values can be misleading right they can mislead you about the status of a given statistic that you are always you know that you are studying. So, whenever we summarize the mean we should as best practice you know data scientists or statisticians, applied economists you know social scientists, and engineers should always also look at the median value ok, to get a full picture.

Second, the comparison between the median with mean signal skewness is something that we have covered before. And hence, the presence of atypical observations in the data is something that we have spent the previous 5-10 minutes thinking about ok.

What this really is helping me to help me do is to build a statistic to infer, whether or not there are you know atypical values in my data right? So, what we say is that if the absolute difference is mean minus median; the absolute value of mean minus median is large enough then a given sequence of spatial data points must be scanned for outliers ok.

So of course, you know we know that the mean is less than the median or the mean is greater than the median both are sort of signals of trouble. But then what is it how far should the mean be from the median that should signal real trouble right?

So, we want to sort of concrete this idea before we are sort of you know before we sort of start scanning for outliers or become worried about outliers in our data right? Even slight asymmetry will cause the difference between the mean and median, but maybe that is not something to worry about so much. So, when you have a very large outlier you are going to have the mean you know pulled away from the median more and more. And after a certain point, it will be a sure shot or statistically, you know significant signal for scanning for outliers.

So, what we really want to do is we want to look at mean minus median and its absolute value why do we look at an absolute value because you know this is exactly the same as you know median minus mean? So, it will not matter which way is it a left-skewed or a right-skewed distribution it will not matter anymore, we just want to look at the distance between the two and figure out what this value should be that should signal trouble. I mean trouble in the sense of outlier values ok.

So, let us move forward and take a look at it, alright? So, what you see on this on your screen now is you know the separate separation of row summaries from column summaries right? So, when we looked at this data when I presented this data to you originally, I had said that you know we can you know we can figure out means and medians in the east-west direction or the north-west north-south direction right.

So, now the rows are indeed the east-west directional summaries and the columns are you know the north-south directional summaries. So, we look at the data the way it summarized is that you know I have my data and it is in its shape; in its shape, you know exactly you know laid out here. So, it is the same layout what that means is that you know the top row which has 5 you know that have crossed in the row are basically representing 8.59, 9.00, 11.86, 8.91, and 9.99 right?

The second row has 7 crosses that are basically nothing but representing value 7 you know coal ash percent values in the second row of the data set layout, which is 11.62, 10.91, 8.76, 8.89, 9.10, 7.62, and 9.65 right?

And for each row when we say row summaries what we are doing is for each row we are documenting a median value which is represented by the circle right. And a mean value which is you know documented by an asterisk right? So, wherever I see the mean is greater than the median that should you know signal right skew that is there is some value that is quite high on the right-hand side of the distribution. So, let me look at the first row I have 8.59, 9.00, 11.86, 8.91, and 9.99.

Quite clearly this value which is closer to 12 is three points higher than most of the values are two points higher than 9.99 you know which is quite a bit if you think about it right. So, three points higher than 9.99 is you know 33 percent higher right two points higher than 10 is 20 percent higher. So, 11.86 indeed is a value which is quite you know peculiar or different and

quite large, given that it's right beside these values on the same row you know where we are sampling coal.

You know keep in mind we are sampling coal, so it is earth. Do you know how different will it be if you are walking on row number 23 and you walk on the first two values they come out to be 12 to be 9 closer to 9? And the third value that you walk on to is 12, then the fourth value that you walk onto is again lower than you know below 9.

So, this sort of signals that there is some peculiarity; seemingly, some peculiarity about 11.86. Similarly, we will go from row to row to row and for each row, we are going to document the mean and the median. Wherever I see mean and median to be very very close those represent symmetric distribution, you know places where I probably should not worry too much it probably seems like there is no trouble right there is no worry right?

Places, where we have very large differences, are places that we should know we should start to worry about right. We will do the same thing in the north-south direction again if I come back and look at you know I am going to start with this particular column it only has one value. So, no surprises mean is exactly equal to the mean; a mean is exactly equal to the median right, so perfect ok.

Let us go back to the problematic value where I had 11.86 which is column number 3 from the right ok let us look at column number 3 from the right. So, I am looking at yeah I am looking at a large difference between the mean and the median. And if I look at this column as well, I have 11.86 which is my first step in the south direction starting from the north. And as I take my second step you know I go down to 7.62 which is quite a fall if you think about it right which is almost you know a 40 percent fall.

Then I go to the next step, I am at the same level similar level, 7.61 same level slightly higher high quite high 9.58. Come back to the same level 8.64 and then 7.83 right? So, this column seems peculiar and seemingly problematic. So, we should now sort of we are now starting to sort of sense some outlier behavior by 11.86 right?

So, we are going to do this for each of the rows right now we cannot be looking at each median and median and mean and each row and each column one by one and it is very time-consuming. Secondly, whether or not the difference between the mean and median is large enough to signal trouble well is highly subjective to the analyst. You know if I am

analyzing these data and I feel that you know the distance this big you know is large enough, then it is highly subjective.

Now, when we do science we cannot rely on the subjectivity of the analyst right we have to come up with an objective measure, which can you know provide us a somewhat definitive understanding of whether or not you know this much distance to say signals trouble ok. So, that is our basic quest going forward.

So, when we do that you know we are going to develop a theory to be able to figure out how much difference between mean and medium; the median is a large enough difference to signal outlier values. So, let us say we have a sequence of data y_1, y_2 all the way through y_n which is independently and identically distributed with pdf f of y with parameters mu and sigma squared quite likely a normal distribution right? I mean that if it is a Gaussian that is why we have a mean mu and you know a variance sigma squared.

The sample mean is denoted as a Y bar and the sample median is denoted as a Y tilde we can write an approximate relationship between mean and median right. So, Cressie sort of provides you know as I said this is all adapted from a chapter, chapter 2 of Cressie's book. So, you can go back and look at this relationship, but this relationship really provides an analytical link between the median value Y tilde.

And the mean value Y bar which we will see in a minute, says Y tilde equals the population mean mu plus 1 over n summation I equals 1 to n sign sgn means the sign of Y_i minus mu. So, I am not really documenting Y_i minus mu, but just the sign of it, right? So, if Y_i is greater than mu I just have a one-it sign operator simply decode or code z into 1. If they are exactly the same it's coded as to be 0, if Y is less than mu I simply have a minus 1. It does not matter what the quanta of Y_i minus mu is.

And in the denominator, I have twice the frequency of the level of mu that I am you know talking about the mean the occurrence of mean in this distribution right ok now we also know. So, we can write the mean as just the population mean Y bar sample mean is equal to the population mean plus 1 over n summation i equals 1 Y_i minus mu. Why is that? Well because you know the summation of deviations from the mean is 0 ok.

So, you know in your basic statistics course you must have learned that the summation of deviations from the mean in any given sample will turn out to be 0 it has to be equal to 0, ok.

So, the Y bar is simply equal to mu we are simply writing this term to be able to create a link between the Y bar and the Y tilde. Then we take the difference between a mean and median right that is what we are after anyway mean minus median right.

So, this is the difference between mean and median and we can write it approximately as 1 over n summation i equals 1 to n the deviation from mean minus the sign of deviation from mean normalized by the frequency of the mean the population mean itself ok. This provides us a measurable metric of Y bar minus Y tilde what does it depend on right? It depends on whether it is a function of the size of the data, it is a function of the population mean, it is a function of the frequency of the population mean right and it is; obviously, a function of the y is right.

So, it depends on these parameters, so it depends on the population you know distribution parameters and the size of the sample, and the values that you are attaining in the sample. So, this difference will is very sample-specific that is the point that I am trying to make here ok alright.

So, moving forward you know when Y_i minus mu is Gaussian, Gaussian means normally distributed right. So, if Y_i is normally distributed Y_i minus mu will also be normally distributed because mu is simply a constant right. So, it is my Y_i minus mu will simply be following the distribution of Y_i with 0 mean and the variance of exactly equal to Y_i.

We can write the variance of the Y bar minus the Y tilde, right? So, now you know remember in the previous slide we wrote down the mathematical measure for Y bar minus Y tilde. Here we are writing the variance of the Y bar minus the Y tilde. Remember, we said that if you have a sequence of random variables the mean of those random variables or realization is also a random variable.

And it turns out, so is the median. So, if the mean is a random variable the median is the random variable their difference will also be a random variable. For every random variable, we have a variance measure that provides us with the error in estimating that value. So, this difference has a variance value which can be written as sigma squared over n times this factor 0.57, which provides us with a statistic u which is nothing but Y bar minus Y tilde divided by 0.57 into sigma over root n right.

So, this 0.75 sigma over root n is nothing but the square root of sigma squared over n times 0.57 right, which is nothing but the square root of the variance of Y bar minus Y tilde. So, what we are saying is we are defining this statistic u as Y bar minus Y tilde over square root of the variance of Y bar minus Y tilde ok. So, this is a statistic we have a data set we have the mean from the data, you have the median from the data you can calculate the variance from the data right?

The variance of Y_i minus Y bar because you have sigma squared can come from the data n is just the size of the data set and 0.57 is just a constant. So, u is indeed going to be a number that you are going to be able to back out from the data. So, if I go back; if I go back two steps from here what I should be able to do is whenever I have a mean and a median for the row, I will also have a Y bar minus Y tilde and I will also have a u for the row.

So, I am going to have $u_{i's}$ where i is just a representation of rows and similarly I am also going to have a Y bar minus Y tilde for each column and I am going to also have $u_{j's}$ for different columns ok. So, I am going to be able to build this statistic that is going to signal trouble when or not the Y bar minus Y tilde is a large enough value ok. So, that is what we are trying to get at.

So, we are now looking at u is being characterized or summarized for each row and each column. So, I do not have to go back and look at each of these values you know visualize whether the mean is very much farther away from the median or not and so on and so forth. I can just evaluate a u and it will summarize for me the statistical difference between you know these values.

Now, sigma squared I mean I said that you can back it out from the data it is just 20 or 27 into inter quartile range the interquartile range is just a difference between the 75^{th} percentile and the 25^{th} percentile how to get them we will just order the data from smallest to largest the 25^{th} value in you know 20 when you are the 25^{th} step. That is the 25^{th} percentile when you have the 75^{th} step it is the 75 percentile right?

So, the 25^{th} percentile will have 25 percent of data that are below this value right and the 75^{th} percentile is 1 where 75 percent of the data are below or lower than this given value when by 75^{th} ok. So, IQR is nothing but you know Y 75 minus Y 25 multiply that by 20 over 27 and you have your sigma hat. Now, notice that u is the standardized mean median difference, and whenever u is going to be greater than 3 a researcher should pay attention to outlier values.

So, what I have done is that I have defined u to be Y bar minus Y tilde over the square root of Y bar minus Y tilde. And, whenever I find the modulus of u which is nothing but the modulus of Y bar minus Y tilde over the square root of the variance of Y bar minus Y tilde ok. The denominator is obviously, always positive the numerator may be positive or negative we are taking an absolute value.

So, it will always be a positive number that we are into look at. Whenever this is greater than three we should be worried about outlier values or troublesome values right, we should start to scan for them more closely in the data ok. Now the point is that you know what it does is that if I go back to my row and column summaries, I can simply you know look at the data and I can just figure out u and then mark out the rows and columns where u is greater than 3 or somewhere around that value right. So, that is what we are going to do next.

So, we again have our data we are going to figure out u for the 23rd row starting from u for the first row and we are going to figure out u for the column first column all the way u to the 16th you know 16th column. Sorry, it is going to be here alright this is probably the 14th column here right? So, we are going to figure out these u values in a table and see whether or not they are greater than 3, ok.

So, let us do that. So, here again coming from Cressie's book we have the standardized mean median difference which is the standardized mean median difference basically is Y tilde. Oh sorry, Y bar minus Y tilde mean minus median divided by variance of Y bar minus Y tilde square root right this difference which is just u is now outlined here.

So, in row 1 the u value is minus 1.54 much lower than 3 or minus 3 right? So, the absolute value is 1.54 much lower than 3 no problem ok, points 0.40 no problem, 6.12 huge problem. We should be extremely worried about this particular row. 4 minus 0.5 4 5 no problem minus 0.35 no problem 2.01 ok 0.56, 0.07, 0.63, minus 0.18 2.12 0.8, 0.46, 0.11 point something 0.61 points something 0.18, 0.35 and so on and so forth.

So, I have just circled the values that I want to be careful about. So, we have 6.12, 63.17 I have also circled 2.87 because it is quite closer to 3 and also 2.47 because you know I think you know why not just check it out. So, what I want to now look at in my data set I will go back and look at rows 3 and 23 and columns 5 and 12. Where these 3 and 12, row 3 and column 12 are clearly problematic and you know columns 5 and 23 let us pay attention to them ok.

So, let us go and make sure we remember these values. So, I am going to look at rows 3 and 23 and columns I remember 12 it was I think 5 and 12 yes, so 5 and 12. So, rows 3 and 23 ok. So, row 3 is right here I go through row 3, I have 9.64, 9.52, and 10.06 a higher value of 12.65, ok, a candidate for an outlier I should use my green color ok.

So, 12.65 is a problematic value because it is lying around this region which is you know in this row everything is 9 again coal existence seems, how come you had such a high value and then a drop right after that right ok. So, my row 3 and row 23 have my 11.86 right, this is the value that we talked about right? I mean if you go east to west or towards south from this value 11.86 sort of stands out right and that is also predicted by my, you know, u variable ok.

Columns 5 and 12; column 5 again has this very high value of 17.61 and it has a high value of 12.80, I am just going to circle them, and I am going to I am flagging outliers. So, that I can look at means and medians etcetera what if I were to drop these values, how will the mean change, how will the variance change will they be robust or not? Column twelve is again we have you know 11.86.

So, 11.86 turns out to be a value which if we were to just look at the stem and leaf plot we might not found find this value to be very problematic right? But as soon as we move on to sort of you know looking at this more sophisticated mean minus median-based, you know, u statistic this value which might not otherwise pop out right turns out to be a candidate for an outlier. This is if this points out a very important difference between spatial data and non-spatial data right?

In space, we are comparing the values of our neighbors and saying you know look we are sampling coal. How different are you going to be at a particular location, can you be really different? You can have a gradient where the coal ash percentage is slowly rising toward the north or slowly rising towards the south or towards the east or towards the west. And you know values on the eastern side are greater than values on the western side it can be a trend that is a spatial heterogeneity measure.

But in a locality can you really expect to see a very different you know coal sample than its neighbors? Well, probably not and that is the concept that this u statistic is mobilizing or using to come up to signal or guide us as researchers or statisticians or data analysts to this value of 11.86 which does not turn out to be you know a value that you want to sort of leave out without attention ok. So, let us move forward ok.

So, now we are going to look at the third you know methodology for detecting outliers and this is this methodology depends on the belief of again on the belief of you know local stationarity. So, when I say local stationarity, I am basically saying a thing that you know in a locality, I expect things to behave normally with respect to each other. And if I see abnormal behavior in a neighborhood well probably I should be becoming more you know more attendant to you know such a phenomenon right?

So, what I am looking at now, you know, what I am going to now propose as a device to do that is called a bivariate scatter plot, between Z at location s and Z at location s plus h e. Now in spatial data, every data point is indexed by a location right. So, Z of s is simply you know the value at location s. So, I can go to any location and call it s and the value that I see there that is the coal ash percentage is going to be Z of s right sorry about the ok.

So, the value here at location s is Z of s alright. Now, what is Z s plus h e? Now e is called a unit length vector with a given direction. So, e is a vector it has a unit length, so it is a unit long right? When you multiply e with h you are going to move h units in a given direction and this direction information is contained within the vector e right.

So, the unit length vector is the purpose or you know the reason that we sort of use this notation of you know unit length vector is to contain the direction information and which we want to move right. h on the other hand is you know is the size of the hop that I want to take from a given location s right, so h is said to be 1. So, I am going to move one step at a time right?

So, e will tell me ok you know if I am standing at s I want to move one step in you know left so that I can get to s plus h e on the immediate cross on the left-right. If I go one step you know south or one step you know towards the downward direction I will have my cross that I go which will become s plus e, e now is the south southern direction right or downward direction from the s one step in that direction right.

So, what we are going to do now is that we are going to create scatter plots in pairs of values where I go to it location I take one step one hop, I record that location and then I create a pair of these things right. So, I am creating a pair of Zs and Zs plus h e and then I am going to now present a scatter plot for these things right?

So, pay attention to what I just said you know I am going to have Z s its going to be my X direction. So, s is in the X direction right X axis, s plus e is in the y direction, so I am moving northward right. So, I am basically starting at each value in my data, right? So, let me go back to my data set I am going to start at 10.59 which is my Z of s and I am going to move one step further to Z s plus 1 right?

And then you know I am going to sort of have 10.59 on the X axis and 9.59 on the Y axis and I am going to point out the coordinate right. So, let us do that. So, I have my 10.5 somewhere around here and my 9.5 somewhere around 9 point something here. So, this is you know approximate sort of a pair of Z s and Z s plus e. Each point on this scatter plot is a representation of neighborhood payers when we move in the east-west direction ok.

So, we are moving in the east-west direction. So, basically in that data set you know if I go back to my data set I am moving from 10.59 to 10.43. So, I am going to be looking at a pair between 10.59 and 10.43 right? But the ultimate purpose is that when I look at a value right and I just step one step further from that value in any direction. Local stationarity the belief of local stationarity basically suggests that you know I shouldn't be looking at a very different value in the locality right?

Again it is coal lying under the ground, how different can you be can the value or coal ash percentage can you get if you just walk one step forward towards, the eastward direction or towards the westward direction towards the north direction or the south direction? It is just one hop, right? So, whenever you know we see values that are standing out from this bivariate scatter plot right that is a signal of outliers. For example, if I get a value of you know when I stand at around 11 and I get a value of around 18 as the next step when I move one step in the east-west direction, I should be worried I am looking at something very bizarre right?

If I go back to my data set again you know I am basically starting out at 10.87 or 10.82 and ending up onto the next hop being such a large value of 17.61 right? So, that signals trouble. So, in this bivariate scatter plot whenever I see outstanding values which are going away from a cluster right a cluster that sort of represents a Gaussian right, a Gaussian behavior like a bell curve.

There is high density in the middle and then towards the periphery, the density will die out. But if you see values that are much further away from a perceived periphery, then you should mark them as outliers right? We can do this in the east-west direction.

We can also do this in the north-south direction, right? So, in the north-south direction again I am seeing the same you know pattern there is a high density in the middle and as I move out in any direction the density sort of has to die out. So, there is some Gaussian behavior going on in every direction from the core of this bivariate scatter plot. So, I can sort of draw a peripheral boundary, but some points seem to be sort of off this peripheral boundary right? So, this is the third methodology to look at scatter plots between Z of at location s and Z at location s plus e.

And these scatter plots provide us a signal of whether or not we should be worried about you know outlier values ok.

So, we have seen three different methods of detecting outliers. So, it is a 3, I am going to call them 3 devices to detect outlier values right? The 1st device was the I am just going to call it a histogram we looked at a stem and a leaf plot. The 2nd device was this u statistic which was based on the difference between the mean and the median and we did this row-wise and column-wise. 3rd is a bivariate scatter plot which is sort of testing the assumption whether or not the data are locally stationary right.

So, we mobilize this concept of local stationarity in the data we will spend you know we will study stationarity in detail. But you get the idea of local stationarity that we are ultimately looking at coal data. And if you are looking at coal data you know we might we can expect some kind of normal behavior and as soon as we see something abnormal through any of these devices we mark it for trouble ok.

So, here are these values that I have you know circled in blue and green and you know they are basically you know a union set union of the outliers that we detected through these 3 different you know exercises. I mean there is this 12.65 which you know I actually detected during the course of this lecture. But I guess you know these 1, 2, 3, 4, 5, 6, 7 and 8 values are the ones that we should be quite worried about right?

Going forward when we analyze these data, whether we are predicting you know we are predicting in space values that we have not sampled right or we are simply evaluating the

mean of coal impurities the average coal impurity in different directions in different locations sub-regions in this region, we should be very careful about these values.

What is special in this analysis that has come from this you know second u statistic is that you know the value 11.86 is very similar to values like let us say 11.62. But 11.62 does not turn out to be an outlier through all these 3 different you know devices that we evaluated right? So, it is not; it is not always trivial straightforward, or something that you can simply with your naked eye just go in and look at the data and you will be like oh here is an outlier I can look at it. Oh well, you can do that that is a good starting point and we should always do that right as a starting point as data analysts.

But we should probably you know also go one step further and work with formal devices or formal measures which will provide us an understanding of abnormal behavior in spatial data ok.

So, that is it for this part of our lecture as the next part, I am going to look at a case study, where we are going to move away from this textbook data set and look at a real-world data set for groundwater levels in India. And we will see how you know when we move from a textbook data set to a real-world data set to know some of the interventions that are necessary before we can analyze this data set.

So, look forward to having you again in the next lecture.