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

Lecture - 10B
Exploratory Spatial Data Analysis: Example of Groundwater Data

Hello everyone, welcome back to lecture 10, and this will be the last part of lecture 10 where we will look at a case study for Exploratory Spatial Data Analysis. In the past you know in the last couple of lectures what we did was we started with you know a coal mine sample data set, where you know we had sample locations at you know in a given datum or a given region.

And the idea was that you know everywhere we do sample we see a percentage of coal ash whereas everywhere we do not sample we do not know what is happening beneath the ground. So, we said look I mean ultimately we would like to sort of understand the coal quality for the entire region, but where we start is by summarizing the data or conducting what we call the exploratory spatial data analysis.

Why is it exploratory spatial data analysis and not just exploratory data analysis, because the data themselves are spatial in nature, they are delineated with a latitude and a longitude cart coordinate system or you know it is just something that we saw in the last lecture a row-column two dimensional coordinate system.

You know broadly we looked at three strategies, the first one was histogram, right? We saw an earlier version of the histogram which we called the stem and leaf plot right and idea the major idea of you know conducting exploratory data analysis or you know this kind of summarizing data that we covered last time was called detecting outliers right? So, we looked at the shape of the histogram, whether is it a skewed distribution, is it a symmetric distribution, and if it is skewed, in what direction that gave us an idea ok.

You know there may be some values that might be quite a large relative to the majority of data or on the other way around you know some values might be very very small relative to the majority of data ok. So, the histogram was the first starting point, but the histogram by itself does not, or did not you know appreciate the spatial nature of the data, right?

So, the second thing that we looked at was this statistic called mean minus median and its absolute value which we defined as something called the U statistic right? And this U statistic was calculated by rows and by columns. So, now, we sort of you know constructed this spatial stat statistic that helped us detect outliers. And we said that look, if mean and median are too far apart then you know we should be worried and we in fact, you know sort of figured that you know if U was greater than 3 and that is the point where an analyst should become worried about outlier values.

The third strategy or major device that we learned last time was based on the idea of local stationarity right. And this basically was is a bivariates scatter plot of observed coal quality or realizations in a local neighborhood right. So, if I observe that 10 percent you know of the coal content is ash at location A if I go one step further in either direction either North or South, East or West I should expect to see a similar value to 10 percent cold right.

If I start seeing something like 20 percent all of a sudden then you know that should sort of you know kind of, provide me some sort of alert signal that you know maybe I am looking at an outlier value in this domain or in this small local region right.

So, today what we are going to do is that we are going to take this machinery that Cressie provided us to conduct exploratory spatial data analysis and apply it to the administrative groundwater level data for India right? More specifically we are going to look at the data that are provided by the Central Ground Water Control Board right and we are going to apply the machinery of exploratory spatial data analysis and try and detect outliers through the three devices that we discovered last time ok.

So, let us look at the data and move forward ok. So, I am just showing you a figure of Uttar Pradesh. So, you know you can sort of guess that we are looking at the administrative boundary of Uttar Pradesh state which is in pink color right, and the dots that you see inside the Uttar Pradesh region or Uttar Pradesh state are groundwater monitoring wells installed by the Central Ground Water Control Board ok.

So, now, this Central Ground Water Board that is trying to monitor values you know every year what is happening before monsoon, what is happening after the monsoon, what is happening right after the harvest of the Rabi crop or the Kharif crop, they basically have to go into each of these, you know wells go in and figure out what is the depth of groundwater level at each of these wells.

And whenever you have you know groundwater depleting or the level of groundwater depth going down quite a bit that provides you know the signal to the extended groundwater board that you must intervene to control these declines and in groundwater levels right. If you watch the news or if you are generally interested in social issues, you would know that groundwater depletion is indeed a very serious issue in India today right?

There are multiple agencies you know there is not only CGWB which is a Center Ground Water Board but there is also an Uttar Pradesh Ground Water Board which you know by which is a; which is a body that tries to monitor you know groundwater levels across the state right? So, both the central agencies and the state agencies in India have their network of you know monitoring groundwater levels because this issue is so important right?

Now, in this you know picture there are dots there are these pointers which are giving us the location of groundwater monitoring wells right. So, these are wells meant to these are owned by the agency meant to monitor groundwater levels remember they are not meant to be drawing water for consumption either for domestic use or agricultural use, or industrial use. They are only to monitor the groundwater data ok.

There are color-coded you know for the year 1998 sorry for the type on that slide this is for the year 1998 and you know the green colors are wells where the situation of groundwater depletion is not so alarming right? So, we have you know the darkest dark green is where the groundwater level goes below the ground about 2 and a half meters we are fine.

And then you know as we go further some wells are colored color-coded red where you know the groundwater levels have gone down up to 32.8 units right? So, these are the wells where you know which should provide some kind of you know management intervention from the agency. Now, very low values and very large values are; obviously, you know candidates for you know for detecting outliers. You know that very low values could be an outlier, and very high values could be an outlier right?

Now, if I look at this data right if I visualize this data I see a lot more dark greens and light greens and yellows than I see oranges and reds, you know for sure reds are you know you can count them here like probably there are 10 you know wells in 1998 that showed very high depletion right.

So, if I were to sort of imagine a histogram for this case the minimum value this the well depth can take is 0 and it can go up to 32.8. So, I am going to make it 33 and most of the values are between you know let us say 0 to 6 in the first two categories and there are some which are going up to 10. So, you know 0, 2, let us say 3, 6, some of them are also going to 10 right and very very low ones are going to the other ones right?

So, we have a situation where you know most of the data lie between 0 to 10, and after that, we have some values which are going up to 33 or 32.8 right? So, this is the sort of you know understanding of the distribution of groundwater levels for Uttar Pradesh ok. So, now, it's clear that if I am going to find outlier values they are going to be coming from you know greater than 10 most likely right?

But, we saw last time that we should not we should be careful, should you know of course, whenever I see values greater than 10 or greater than 15 or 18 for that matter that should definitely you know create some kind of you know alert, but in general as well like you know if I see values greater than 10, I should be you know starting to detecting outliers more formally than just looking at you know a histogram or the visual picture.

So, let us move forward. So, we saw last time that in the presses machinery, the data were organized in the row column you know in row-column fashion. Now to do that what I do here is that to the map of Uttar Pradesh where I have many wells scattered around space and the spatial distribution is not consistent. Now, this is not textbook data, I will find a very nice characterization where data will be you know exactly collected in a nice grid cell of rows and columns.

But I have to attempt to organize the data in that fashion so that then I can apply the machinery that Cressie has given me right ok. So, ok we apply this grid cell formulation and we try and organize data in this you know in this format to move forward ok. So, we have to apply an intervention to the given raw data set and then you know to be able to move forward with the textbook methodology that we have seen till now ok.

So, for the purpose of this lecture, we will look at a very small area, we will not look at the entire state because you know it will be a lot of data and hard to sort of you know bring together for the pedagogical purposes. So, we will focus on a small cross-section in you know, and sort of near Central to Eastern Uttar Pradesh.

So, I have seen I have shown you a zoomed-in version of this little cross-section. And what we see here I mean we observe a few things, the first thing that you observe here is that many of the cells have no values and have no monitored values right? So, this is a real-world dataset, if we do not have a value in a cell, well we do not have a value in it in that cell, and we cannot do anything about it right.

Now, Cressie's methodology does not necessarily exclude this kind of situation, if you saw although most of the data were nicely organized in the row-column grid network from the previous lecture many of these grids were empty right? So, we can now even with this situation real world situation we can move forward with Cressie's methodology right.

But another thing that we see here which is much more critical is that some regions in this section seem to be much more highly sampled than others right? For example, you know if I look at you know this region in the green circle there are two regions that I can find which seem to be most more densely monitored by the monitoring agency that is a central converter board than other regions let us say on the East of this cross-section right?

Now, there are many reasons why this can happen right, it is possible that there are some regions that require more intense monitoring right, some regions that are more prone to you know depletion of groundwater, and hence the monitoring agency decides to go in and monitor that region more closely right. Other regions may be you know even remote I mean it is possible that it's a forest patch or it's a water body.

As you know I cannot naturally through due to natural constraints and cannot really monitor, if that is the case I am not going to find too many wells in that region right? If the river Ganges is flowing through a particular region you are not going to find you know monitoring wells in the middle of the river Ganges right?

So, there are natural constraints, there could be you know monitoring requirements that can drive how different regions are monitored you know across space, especially for groundwater. And they may not follow the nice schematic textbook schematic that Cressie had provided us where you had equidistant sort of monitoring sampling locations distributed across a datum for the most part ok.

So, if I come back to this zoomed-in version, you know, I see that there are some cells that have two monitors, two observations and you know I am just going to try and mark these

cells for your benefit here right? Where I am able to see some cells with you know more than one observation filled in. So, you know these cells basically are containing more information than their other you know counterparts right? So, what is the consequence of this kind of you know sampling methodology?

So, we are going to sort of learn about something called sampling bias in the sense that some regions are more densely monitored or sampled than others right? And this can possibly you know introduce bias to our samples and before we go ahead and use these data for summaries summary statistics for doing you know calculating the u statistic that is the median minus mean minus median absolute value, we should do what is called as declustering of this data.

That is to be able to make comparisons which are apple-to-apple comparisons right, just because you have three monitoring locations in a given cell you know should not favor you in some way in terms of the value that you are able to provide versus other places where you have only one observation or no observation at all right. How do we create a balance, so that when we are you know when we are comparing the means of values in a given cell with two observations versus one observation remains an apple-to-apple comparison rather than apple to oranges comparison ok?

So, let me first sort of go through this idea of sampling bias and cell declustering, and then we will apply this to this data and then move forward to you know conducting the analysis. So, the idea is that you know I cannot move forward with the analysis till I have intervened and applied cell declustering to a real-world data set right? So, this is an extra intervention that this real-world data set demands.

So, let us define this situation, at times while using real-world data sets we encounter situations where sampling density is higher at some locations or localities or regions than others ok. And of course, you know there might be multiple reasons I mean there could be the reason could be access or convenience right, just because it is convenient for me I can sample that requires location as a researcher.

So, it is a subjective sort of bias that might enter through the person who is sampling data right. So, there could be this kind of bias. They could be in the case of you know groundwater you know they could be things like a crisis, community attention, attention and even media can drive sampling of groundwater levels right? So, you know especially if you have a lot of

you know attention from the community because the groundwater levels are declining and we will not get water in the future.

And it is possible that the monitoring agencies can intervene and start to monitor those areas more intensely. Now, what that will do if you try to visualize that situation you know that last example that I gave you if you try to visualize that you know what will happen is that there is going to be a very large sort of density of you know samples in an area where the water levels have declined, relative to ones where they may not have you know declined as much right.

So, you can have a situation where you have a region, where such that high values are sampled more vigorously than low values right? In that scenario, if I were to take a sample mean if I were to just take a mean of values you know through the datum without this considering why some areas are more densely sampled than others I am going to under weigh the regions which may not be under crisis and overweigh regions which are undergoing depletion at a faster rate.

And the overall picture will come out seeming like; there is so much crisis right whereas, what we have done is that you know where we have you know in this edge region where we have so many sample locations, you know if I add one more I am not really adding any more information, but relative to that if I were to add this monitoring well you know these two monitoring wells in the low region I will get much more information right.

So, that is the crux of this disproportionate sampling you know distribution in space right, and the kind of bias that you know this issue can bring right? So, here; obviously, you know H values are overrepresented; are overrepresented and L values are underrepresented correct right? So, if I take a sample mean if I just take a mean without this without being cognizant of this disproportionate issue I am going to have a biased understanding of the situation in the real world that I am trying to study.

And it could be anything, it could be development outcomes, it could be education levels, it could be crime rates, it could be you know groundwater levels, as we are looking at it you know it could be deforestation, it could be many many different you know contexts that one can study these issues right. So, this methodology that we are studying is much more general right ok.

So, you know the solution to this problem. So, the solution to this problem of sampling bias where bias is introduced due to how we sample data is called cell declustering ok. What is cell declustering? So, cell declustering for cell understanding let's actually draw a figure which is quite similar to the real-world situation that we are spacing with our dataset.

So, we have a grid cell right where we have sampling locations distributed across these cells right, some of these cells, or most of these cells you can say will have one observation, but some of them can have more than one observation ok. So, I am going to simply try and ok, let us say this is the data set that I have, some of the cells are empty and not represented, some of the cells are represented by one sample location, and some other cells are represented by you know even a situation where you have 6 sampled locations ok.

So, you know we have divided this area into a great grid. So, we have let us say a total of you know capital N cells ok. And the number of data points that are observed in each of these cells can be let us say you know denoted by a letter C right C observed ok C observed right?

Now, what you can do is let us say you have a total of T data points in the entire domain right? So, total T data points in the entire domain are right in the total right? C_0 or let us say not C_0 let us call it C_n. So, I will just give it a notation of small n. Just a second ok, C_n is the number of observations or data points in each cell you know that we are looking at ok.

Then define a weight at each location and let us say S_j right. So, this is cell j, right? So, j^{th} cell I have you know this weight that I am defining you know I could in fact, just call it Wn right Wn and this will be equal to I am going to define this weight as C T over N. So, this is the total number of cells, the total number of observations per cell that I should expect if the data were you know uniformly distributed and this is going to be now weighed by the number of cells that I actually observe ok.

So, this is the ideal this is the actual right, this provides me a weight that if I apply to each observation in this cell right each observation in the cell then I will get a decluttered value ok. So, decluttered data values will be W_n Z_n this is the weight right for cell n and this Z_n is the data observation in cell n. Remember for the cases where you have just one observation this weight will be higher than you know for those cells than the cells which had you know more than one observation.

So, W_n will start to discount the density of observations in highly sampled cells right? So, the fact that you might be over-representing you know H regions will be discounted or normalized by this W_n weight right?

So, using this weight we can then define our summary statistics we can say the declustered mean, what would be a declustered mean will be Z bar equals summation i equals 1 to N W_n times Z_n divided by i equals 1 to N or let us say n equals 1 to N W_n ok. Similarly, we are going to have we can also define.

So, this is the sample mean, I can also define sample variance, let us say this will be S squared equals summation n equals 1 to capital N W_n Z and minus Z bar. Remember Z bar is the declustered mean ok, divided by summation n equals 1 to N W n minus 1, such that summation n equals 1 to N W n should be equal to N ok.

Similarly, you can define sample covariance ok. So, these are the summary stats that you know we typically evaluate with the data right? So, we have again summation n equals 1 to N W_n Z, or here it is x. So, I am going to use x n minus x bar again x bar is the declustered mean right y n minus y bar x bar and y bar have to be equal because they are declustered means for the same region I could in fact, use Z bar instead of these values right over summation n equals 1 to N W_n right.

Now, I have been able to construct you know declustered means mean values or summaries or data-based statistics which will not be impacted by disproportionate sampling in space right? Now, a question that would arise is you know what is a good cell size right, what is a good cell size? Now there is no real theory for what should be the right cell size. So, let us just go back and let me show you.

I mean it is possible that I could have instead of using the cell size as you see them on the screen I might have you know instead of using larger cells right I mean I could have combined 4 cells into 1 and created a grid network that looked much bigger right? So, you know, and if I do that then my problem sort of becomes a little different right, I mean my number of total cells will differ, the number of you knows observations and each cell will differ and so on and so forth.

So, my entire analysis is conditional on the size of the cell that I choose. So, there is no theory for choosing these cell sizes. The best strategy is to visually analyze whether you

know in sparser regions we can capture let us say one datum one data point right and in dense regions, we are still able to you know sort of get into you know individual data points are then you know encompassing a lot of data points at once.

With larger cell sizes what would happen is that I will have some regions which will sort of start to encompass so many data points and most of the other regions will have nothing right. That is why maybe larger data and larger cell sizes may not be optimal for the case that I am working with right?

So, it is really an analyst's judgment of what should be the right cell size, but the point that I am trying to make here which I want to sort of drive home is that when you know work with a real-world data set of the choice of your interest then you know the best practice would be to make sure that you know these you do you decluster these data points otherwise sampling bias will enter your analysis ok.

So, now having understood this we will now move forward and apply Cressie's machinery and try and discover outlier values for our data. So, let us apply Cressie's method machinery here, each cell has now the declustered mean right. So, if you had two observations in a cell you know you have the declustered mean value inserted in this cell right this looks very similar to the x y, you know column, row, you know representation of the Cressie's data that we looked at. In all, in this data set, we have 37 rows and we have 22 columns, right?

So, we have rows that are represented by the x-axis something that we saw in the previous lecture as well and then we have you know the columns on the x-axis right. So, rows on the y-axis and columns on the x-axis. Now, let us look at the make above data. Remember, we looked at the entire UP region data set and there the outlier values were sort of you know greater than 10 or you know we should be alerted we should have been reverted we saw values greater than 10 certainties of greater than we found something greater than 18.

Remember, this is not the entire UP region data set I am looking at a small cross-section right? So, if I am looking at a small cross-section you know I should be sort of cognizant that I cannot apply the same summary stat that I sort of applied or I you know earlier. So, I should do a new stem and leaf plot or a new histogram to get a sense of these cut-off values where I should be you know alerted for outliers.

But still, I can make a start you know I can just look at the data and what I see here is in the first column I see 11.13, 16.84 pretty high values, 1.95 again a very small value, 4.58, 1.97, in the second one I see 12.5, 5.9, 7.02, 9.3, 4.1, 2.2, 8.9 and so on and so forth. The entire mix of data sets seems to me that if I see a value greater than 10, I should still get worried right because I do not see too many large values here right?

So, if what I see is a lot of 3s and 4s and maybe 6s, but there are some you know 12.3, there is a fifth you know 10.5 all of a sudden right? So, those sort of those values sort of should sort of ring a bell that you know maybe something is off here right?

So, let us move forward and let us do these mean and median summaries to get that sense ok. Now, we know that if the mean and medians are far apart that is the first signal that you know there may be outlier values to look at. These are the mean and median summaries across latitudes. So, the latitudes are my rows right, if I go back sorry these are my columns. So, you know I am looking at longitudes which are 37 rows and 22 columns.

So, if I have data you know ok. So, it's across rows, and it's across longitudes not attitudes sorry about that right? And my median is represented by a green circle whereas, the mean is represented by you know a cross of red color. There are areas where the mean and median are quite close to each other, those are non-troublesome areas, but there are other areas which I should start to worry about like, for example, row number 6, row number 29, you know row number may be row number 2, row number 3 and so on and so forth right.

Wherever I have these issues you know I will have to go back and look at my data again for outlier values ok. So, I am going to say I am going to mark rows let us see 3, 6, 29, 31 right where else do we see a very large difference between the two maybe here we see a very large difference at 33 ok. So, we are going to look at 3, 6, 29, 31, and 33 ok, we will try to remember that when we get there well just do a recall ok.

Now, this is across columns we have 22 columns, now again mean is in red a dead dot, the median is a green dot and here you know there are many columns where things look ok you know things look alright to me right, but there are many other columns where things do not look ok. So, for example, column number 2, column number 8, column number 9, and column number 13. So, I am going to write that down in columns 2, 8, 9, 13, and 20, right?

I should probably also check all these other columns right I mean they are not; they are not exactly quite small. So, you have 14, 15, 17 alright, and maybe this one as well number 6 ok. So, I am going to make a case for you know I am going to recall them when I go back to the data set and I am going to look at these columns again carefully ok.

I am back to my data set. So, here what I am providing you is u statistic for columns ok. So, for columns, if I go back you know I have what seemed problematic was 2, 8, 9, 13, 20, 14, 15, 17. So, 1, 2, 3, 4, 5, 6, 7, 8, and 9 columns look problematic visually. Let us go back and look at u statistics and see which ones are actually problematic.

It turns out most of the columns are fine there are only two columns and maybe one more right here which is 2.78 which I should probably pay attention to the right. So, among these 9, 8, 16, and 18 are problematic right? So, in my visual analysis, there is 8, there is 16, I do not even have 16, I do not even have 18. So, maybe you know. So, this one was more problematic than the one that I picked right.

So, this is the reason why you should do you know a summary statistic right or use the U statistic because u statistic is not about just the difference, but also the variance of the difference. Especially in the case of row 16 right, if I go back and if I look at row 16 I have a value of 2.78 right U is y bar minus y tilde over the variance of y bar minus y tilde ok. Mean minus median over the variance of mean minus median right or sorry maybe the standard deviation is the square root of the variance right?

Now, if I look at 15 and 16 if I go back and look at 16 and I look at 15 and look at 17 the value of y bar minus y tilde seems to be higher for both 15 and 17 than it does for 16 right, but u statistic controls for the variance as well. So, it turns out that although 15 and 17 rows 15 and 17 have higher numerators they also have a very high denominator.

Whereas in the case of 16, the numerator is small, but the denominator is proportionately even smaller which is why this row in this column turns out to be more problematic. Let us look at these columns where U is greater than 3 right? So, this is U greater than 3, U greater than 3, U is less than 3, but close to 3. So, I am just an analyst I am a little bit more conservative and I want to be careful before I reject something ok.

So, here I have 2.72, 2.2, 0.68, 8.03, 1.44, 12.71, 2.05, 2.56, 1.78, 4.55. I would circle the high values of 8 and 13 or 12.7 as potential outliers ok. Similarly, I can go to row 18 where

the U is definitely greater than 3 by its probably the highest. So, I have 2.3, 3.5, 1.8, 2.42, 3.71, 0.8, 2.23, 3.55, and 10.5. So, I arrived at a high value working through this column 2.3 and 3.17 which is another candidate for an outlier right?

Now, let us look at this little you know a problematic seemingly problematic row column which is 16, 1.6, 3.9, 4.8, 1.4, 0.53, 1.1, 0.842, 1.47, 1.15, 12.05. Indeed I am not able to find a value that I should probably be concerned about. So, you know U less than 3 seems to work for me alright. So, I am going to move forward with these 3 candidates of outlier values, before I move forward I am going to now give you homework where you should you know conduct the mean minus median analysis for rows and point out the outlier candidates, point out the candidates for outlier values ok.

Obviously, this is not going to be a created homework, but this is very important for you to practice. Look you have the data in front of you on the screen, you have the methodology just apply for practice right, this is practice ok.

So, I am going to move forward and I am going to look at these plots right these cat these bivariate scatter plots between Z_s and Z_s plus e right? Now, these are across columns. So, this is not to South. So, I am going to stand on a location, I am going to go one step North right and create a counterpart of Z_s as that Z_s plus e right.

Here now remember when we think about it we look we try to find a core, we try to imagine a normal distribution in space, and then think about the outlier values. It seems like the core is somewhere here or maybe like even downward and you know the distribution is sort of you know going to sort of fall down from 0 further on right.

So, I think the values that are outside far apart from this cluster of data are good candidates for you know outlier values right? So, I should go back and check all of these Z s es and see if they fit the bill to be you know outlier value ok alright ok. This will be here and so on and so forth.

So, this is now ok. So, the previous one was West to East, and this one is North to South, sorry I kind of gave you a different understanding earlier. So, the previous one was West to East and not North-South, but this one is South ok alright?

So, that is about it. So, I have not really given you the outlier values I am still leaving it up to you, to figure it out by yourself and complete this process and come up with the candidate outlier values that you should be worried about when you report the mean of this data, the variance of this data, the standard deviation of this data and so forth so on and so forth ok.

Next, in the next lecture, we are going to cover this concept of spatial stationarity right? spatial stationarity is a very critical concept of spatial statistics and if we are to define a mean, variance, or anything about a data set we should first resolve spatial stationarity alright. So, look forward to having you in the next lecture on spatial stationarity.

Thank you very much for your attention.