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

Lecture - 13B
The experimental variogram on an irregular lattice

So, welcome back to another part or another section of lecture 13. In this, we will start by looking at an irregular lattice to calculate a variogram or a semi-variogram and calculate an experimental variogram to be more precise, right? We saw how to do it for a regular lattice in the previous part of this lecture.

So, what you have on the screen here is an administrative groundwater data set that you have, we have seen, we have worked with when we did exploratory data analysis spatial data analysis, right?

And these data are an irregular lattice that is to say that if I were to divide the domain of interest which is the state of Uttar Pradesh into equidistant cells that are sort of you know marked by rows and columns in i and j direction or x and y directions of a certain size. I will not have data that is going to be observed in every cell of you know in the entire domain right?

To see an example, we can look at any vacant space within the spatial domain of interest. Whenever we see a space that is vacant for example, the square on the screen represents the region that remains unsampled in the year 1998. And that is not surprising, you cannot possibly put wells, these are wells that are going to have to be dug into the ground to a certain depth.

So, you cannot have at day zero all those places being sampled, what is encouraging is there is quite an intensive sampling even in 1998 when you know these data sets were first made available right? Now, when if you sort of see what happens if you have a row-column representation that is data that are contained in cells in the entire region.

In this cross-section that we are looking at, there will be no data and what is the consequence of that? The consequence is that if we were to sort of transport the idea of collecting this neighborhood set an h by starting from a tail and going up to a head in a strict direction as you know encapsulated in the definition of a spatial lag vector h. Then we may not find any

corresponding you know pairing observation from this tail right; so, we will have a vacant observation.

So, for a lot of locations while for some locations we are going to find such pairs there going to be a lot of locations when where we are not going to be able to find any of these pairs right. So, what do we do then right this is very important because see most of the real-world data sets are going to be similar to the one in front of our screen right now right?

In most places when people monitor air pollution levels, they are not going to be able to put these air pollution monitors in nice looking grid format. So, that you have a regular lattice, you are going to encounter irregular lattices when you are working with these real-world data sets ok.

So, let us see what does what is the workaround. The workaround is that when you have a tail to a head you know a lag vector that you can scan around every part of this data set, you do not make its direction very strict. You allow for the fact that you will be able to keep the tail fixed you will be able to scan a little bit angular you know in an angular direction from the strict direction h which is let us say west to east.

Not only that you also want a possibility that maybe you do not have in the strict east-west direction you do not have any value observed at the head. But if I were to sort of extend the head to be slightly larger that is h plus delta then I might have an observation in space. And that is now that gives rise to a strategy of observing data you know to create this neighborhood set h.

That is creating unique pairs that are you know separated by h right? So, what we do is instead of taking our scanner you know strictly like a line we basically take a v-shaped angular tolerance which allows me to look at the scan for you know spatial lags in a larger area. Of course, we cannot make this too big; we cannot make the scanner too big because then we will be going in some other direction altogether, and we do not want to do that.

We just want some tolerance, so, when we go from lag 2, we are not so strict, we are also able to sort of you know look for values at an angular tolerance. We are also able, we also want to be able to look for values slightly further away from lag 2 which is the h that we are working with, so, this is h and this is h plus delta. Of course, we do not want this delta to be too big, it's small enough, and it is a little tolerance in how far apart are we considering the lag h.

Of course, you know the tolerance could also be on the negative side, I mean in a smaller distance you might find a little value and you are like ok fine, I am going to take it ok. So, we sort of you know instead of a strict line you know a representation of a lag vector h, what we do is that we come up with a more envelope you know the formulation of this lag vector. It is a lag envelope rather which has angular tolerance and it has length tolerance in both positive and negative directions ok.

So, having said that let's apply this idea to our data set. Now, my irregular lattice is given at the top and the bottom from left to right. I have three different representations of the lag vectors. The size of the lag vector is the same, so, we can say h is fixed for all three directions right? When I say three, I am pointing to the three figures on the bottom half of this slide, right? what you see is that h is different due to its direction.

In the case of the first figure, we are looking at a north-south representation of spatial lag. In the second it is the east-west or rather west to east whichever you prefer and in the third one we are looking at a southeastern direction so far as you know the spatial lag representation is concerned.

In the north-south, if indeed my lag vector was so sort of fixed to be from north to south you know all the lag representations in blue should have been parallel to each other which is not the case right? Some of them indeed will look like to be going exactly in the north-side direction; for example, the one that I have marked on your screen, but some of them will take advantage of an angular tolerance.

Because I did not find a head exactly in the south let me just go slightly right in the direction and find one for myself right? so, this search panel allows you to collect data with that kind of tolerance. It is also possible that some of these sticks from north to south are slightly larger or smaller than the exact value of fixed h that we started with.

Similar is the case for the East-West direction, you do not have an exact parallel. You have these in inclined you know h vector representations which are taking advantage of the angular tolerance in our search for the spatial lags or our scanning exercise for spatial lags and the same for the third figure. So, you can stop here for a minute and just visualize how to go about scanning or searching for spatial lags when you are given an irregular lattice.

On the next slide what I have for you is what I am calling the semi-variogram cloud. A semivariogram cloud first of all I am saying is built on ArcGIS. So, interestingly I am now also introducing software on which these things are calculated. It can look rather intimidating that I have to go and scan all the different lags and then collect those pairs.

And then I have to sort of you know then calculate the mean squared value the difference squared values for each lag and then sum them and take a mean, right? well, the software will do it for you. Every data point, every data point that you see on this graph is representing a given lag vector h; for example, the 1, which I have marked in blue represents a lag of about 0.3 units right; so, h is 0.3.

For this h, I had Z of S and it is Z of S minus S plus h which is 0.3. Let us say, it is an approximation, ok, do not take it literally, I take them and I square them and I am looking at a semi-variogram. So, I divide this value by 2 and that is the value that I observe here as gamma which is around 4.5 you know into 10 to the power minus 2.

So, this is my gamma h which is given as Z S minus Z S plus 0.3 in the direction of interest. The direction is east-west or west-to-east, right? you can see clearly and I am able to sort of get this point. So, for this 0.3 units, you can see clearly that there are many many pairs, there are many many pairs of interest right? So, I might not have been able to mark all of them, but all of these pairs collected together is my neighborhood set N h, the neighborhood set N 0.3 n with 0.3 in parenthesis right?

So, a bag that collects all these values; a bag that collects all these values right? A set collection of all unique pairs at h equals 0.3 is denoted as N of 0.3 or N of h equals 0.3. And the modulus of this N 0.3 is nothing but the count of such values. So, what will be the count of these values? When we simply count all the yellow crosses along the stick that is h equals 0.3.

So, this is semi variogram cloud it is like a scatter plot of data right? The very important thing that we see here is that till now we have whenever we define the experimental variogram or calculation of the variogram we kept h fixed, what we have done here is that we have variable h ok.

When I talk about variable h, I am talking about you know process where I am first collecting the unique sample pairs of data or data pairs that are separated by h. I am varying, I am

keeping a fixed h value, but then I am varying h direction. Second I can vary the h value while I have a fixed h direction, for all the permutations and combinations that I get a given value of h on this semi-variogram cloud, I am only using the distance metric.

So, I am plotting all the directions with h equals 0.3 on this vertical line that represents h equals 3. Now, you know; so, we have moved one step further, we have calculated 2 gamma h for all these unique h values right? let us just write this down; so, that it is absolutely clear in your head what we are up to here ok. And then we vary these h values; so, we create 2 gamma h 1, 2 gamma h 2, 2 gamma h 3 keep going let us say you have a total of capital M you know h values for which you can conduct this exercise.

This is the set of all experimental variogram values at different spatial lags right. Now, spatial lags will depend on what the domain size is, what the domain shape is and so on and so forth, whether you are working with a regular lattice, you are working with an irregular lattice, and whatnot. But the point is now I have more than one representation of 2 gamma h, h itself has an index j right? j provides M a representation of how many lag vectors am I collecting the data for, this is all going to be an analyst's choice.

I can have very fine h, I can have very coarse h values that will determine how large the capital M value is. Once I have this set of different 2 gamma h values, what I am going to be able to do is, I am going to be able to have h on the x-axis 2 gamma h or gamma h either the variogram do it does not matter it is just scaling by 2 by a multiple of 2 right?

And then I am just going to put this one value; remember the cloud is different from 2 gamma h, the 2 gamma h is a unique value. The cloud is a collection of all these you know z_i minus z_i plus h squared values ok. So, then you know, we will be able to sort of figure out what these values are at different values of h. So, going back we should be able to make a sense of what these things are, when I said gamma h equals this just be careful, I am not you know comma h is not defined like, this gamma h is a mean of all these unique values.

So, I, should not have used the representation its gamma h. It is just an experimental cloud version of gamma h, right? it is not the exact definition of sample gamma h that we see on this slide or what we have studied ok all right?

So, moving forward; so, we will take a digression now and we will ask a question that is the variogram a resistant statistic? What is a resistance statistic? To always sort of get a sense of

a resistance statistic we can recall the mean and median of a distribution. So, we have seen earlier that the difference between the mean and median is a reflection of whether or not a given distribution of data, has a symmetric you know PDF well.

More than that the distance the more distant mean and median become they are also a signal for outlier values right? The mean value is pulled away by the outlier value in its direction right; so, if I have a left-skewed distribution where the outlier values are sort of towards the right of the distribution, what will happen is that the mean will sort of get pulled in the direction of the outlier right.

Whereas the median is more resilient to it right, to get an example you can simply take a sample of a sequence of numbers from 1 to 10, calculate its mean and it calculates its median. Now, to this sequence add a number 100, again calculate its mean and calculate its median you will see that the median remains resistant to this outlier value 100 or 1000 to this original sequence of 1 to 10 right?

Whereas the mean sort of runs you know in the direction of the outlier. That is why the difference between the mean and median provided us u statistics, remember the exploratory data analysis allowed us to sort of figure out, whether or not we should be worried about outlier values in a given sequence of data. So, the variogram suffers from this issue of outlier values, to see this just realize that if I have zs_i being 1 being value 10 and zs_i plus h being value 11.

Then the contribution to the variogram is exactly 1 unit which is 10 minus 11 the whole square which is just 1. But instead, if in local you know at the lag h of s_i had observed a value you know let us say 1000, then the contribution will become 990 squared which is a huge contribution to the 2 gamma h value. This will pull 2 gamma h to a greater positive value which is then a reflection of lower spatial dependence.

So, having its outlier value will create this misjudgment of lower spatial dependence in the data. To see this, look at the bivariate scatter plot that we have seen many times during this course now. Now, for the value which is let us say 11 you know the corresponding z_s plus h or the value nearby is close to 18 right?

If I were to not remove this and calculate 2 gamma h at location 11 with h_1, it will be pulled in you know quite a bit by this difference of 7 between these values right? If I include these,

the data seem quite scattered and the correlation of the covariance and this data seem to be low. So, the spatial dependence is sort of becoming weaker due to this spread that the outlier values are bringing to this scenario.

If I were to exclude these values and only focus on the values you know that are in the middle then the correlation seems quite high, right? things seem to be moving in a direction closer. And as I sort of keep on excluding the outlier values further, you know I will have a smaller core which will look more and more spatially dependent. So, the variogram by itself is not a resistance statistic, it can lead to a $m_{i's}$ estimation of spatial dependence in the presence of outlier data values.

That is why we conducted an exploitative analysis before we introduced the variogram right? That is why it is very important to exclude the very outlier values before you go on to do space conduct spatial prediction or conduct spatial regression ok. Because in the presence of these outlier values, the covariance structure the spatial covariance structure in your data is going to be messed up. And everything that will follow no matter how sophisticated your video analysis is will be a misestimation ok all right?

So, there is a resistant version of the variogram that is not so popular, well it uses the modulus of the difference which does not allow the penalty to be squared it takes the penalty and square roots it right? And then conducts some power adjustment some normalizing factor adjustment, but it provides us a median analog of a resistant variogram, right? So, Cressie and Hawkins provided this version of a resistant variogram in 1980 right?

So, this is just an understanding of you know when we have the z_i z_{si} minus z_s the whole squared you know values which are the semi-variogram or the variogram cloud. By itself including the outliers, you can see that the distribution is very highly skewed to the left and right. Whereas, if I look at the representation that the resistance statistic is using which is the square root of the modulus or absolute difference between these locally situated values.

You know then the distribution seems much tighter and perhaps closer to what we are used to ok. So, that is the utility that we see experimentally, we can see that even with the groundwater data that is Uttar Pradesh data that is a real-world data set. We see the difference right away when it comes to measuring the spatial dependence of the data ok.

And these very small values will exhibit large spatial dependence that is to be expected you know its groundwater data how much will it change as we move through the space right? It is a geographical structure right beneath our ground, but it is not going to be like tubs or you know walls built together it is a large tub over space right ok alright?

So, let us come back from the digression to this semi-variogram cloud. And now, sort of start to build the experimental variogram which is taking the mean value at each h you know h value taking the mean of this cluster points at each h value that will give us the 2 gamma h representation as we understand it.

So, let us see, how the experimental variogram looks like. Not a surprise the experimental variogram is just a point at each h value. And the point is nothing but the mean of all these you know scatter plot points or the cloud points z_s minus zs plus h squared. So, you sum them divided by the total number of such values at a given h which gives you a 2 gamma h value right.

Now, we have understood a few characteristics of this experimental variogram, you know as a class exercise what you can do is now take a 2-minute pause and locate the nugget, the range, and the sill on this variogram. So, I will come back in 2 minutes pause your video come back in 2 minutes and then I will explain where to locate the range, the sill, and the nugget of this variogram. We studied these with the theoretical variogram all these parameters are the properties of a variogram.

Welcome back; so, to locate the nugget range and sill we simply have to fall back on their definitions, what is a range? The range is the lag distance at which the variogram reaches the sill ok, the sill is the large-scale variation in data remember its C 0 right? So, it's a level of 2 gamma h value when there is no correlation in the data. So, the sill is just the distance between the origin and the vertical distance to the no correlation point is the sill, this is the sigma squared C 0; however, you want to represent it.

At the so, you know the point from the origin on the x-axis that the farther that we have to go to get to this point that this sill is called as the range. The range represents the distance from any given location to which there is some spatial dependence in data, beyond the range, there is no spatial dependence in data.

That means, for our groundwater data set the groundwater level at the tail will provide no information about the groundwater level at the head where the head lies farther away from the range value which is the threshold value after which we have no spatial dependence, no information whatsoever to predict what to expect.

The nugget is a micro-scale variation we know its notation it's C 0 right? On a theoretical variogram, the nugget is something where you know we have C 0 right? So, nugget is the 2 gamma h value when h approaches 0 right; so, this is 2 gamma h when h approaches 0. Clearly, and something that we have discussed earlier that the data are usually we are not able to collect data very very close to 0.

In this case, I need a data point that was at a location where I just moved out from 0. So, I needed data points that were very very close right; so, I needed to where would this 2 gamma h value be just right outside this point location s 0; so, I should be able to do it for all sample points if not sum. Now, we said that usually we do not have these understandings or we cannot realistically collect data.

If we are digging a large enough sort of monitoring well for groundwater, we are doing it at a location you will never see a well right? Besides it you know it is nonsensical right? So, what we do is that we consider this C 0 to be composed of a measurement error; a measurement error, and a white noise; a white noise representation of it.

So, it is predicted from the data rather than being calculated or estimated directly. And the sill is just the large-scale variation in data something that we have discussed in detail now. I hope this makes your understanding of an experimental variogram very very clear you know I sincerely hope so, ok.

So, some couple of last sort of you know ending pointers in this lecture is that by looking at the variogram we saw that you know we started with the bivariate scatter plot which we called the x scatter plot and we said we can get a sense of the variogram value. I am now going to go back and completely close the loop that is we can start with the variogram and predict the variogram plot, and variogram graph. And predict how the co-variogram or the h you know the bivariate scatter plot which is the eight-scatter plot could look like right?

So, consider what we are looking at you know on the left-hand side, let us say we collect our h value here, let us call this h tilde. At this h tilde, you have a large variogram value; so, the 2

gamma h or gamma h which is a semi-variogram value is quite large right? At this level we know there is no spatial dependency data no or little spatial dependence in the data. Hence, the bivariate scatter plot will be scattered to a large extent around the 45-degree line which is a representation that perhaps there is no correlation in these data at that h value.

But what if we were to look at this at a different h let us say let us call it h star ok. At h star what happens is the variogram value or the semi-variogram value is small. So, if the semi-variogram value is small; that means, there is a significant spatial correlation in these data. There is quite a bit of spatial continuity a spatial dependence going on around this value h right?

At the distance h from any given location, I will have a healthy dependence, that healthy dependence will mean that the data will be scattered more tightly around the 45-degree line right? Why? Because if all these scats and all these data points in the edge scatter plot were to sort of were to be located on the 45-degree line we have a correlation of 1 which is a perfect correlation.

That is the place where you know the value of you know if at any h star or h double star the value of gamma h drops to 0 that is a perfect correlation, I have never observed such a situation with real-world data sets, but theoretically that what it is right. So, when we started this lecture, we said we can sort of you know look at the h bivariate scatter plot which is coming from the local stationarity idea something that we have seen earlier, and start to predict what a variogram value could be you know experimental variogram look like.

Now, I am saying we can go the other way round to, which is it should make intuitive sense right?

Finally, spatial contiguity you know which is a smoothness in the data set overall large scale spatial continuity means how spatially dependent values are in their local proximity. Now, if we look at the first figure it is a course image, right? I mean it is an image let us say you pick a picture from your camera you keep zooming and you come to a very coarse pixelated understanding of the world.

Whereas, you can have a smooth image which is let us say, let us call this image 3, and from 1 to 3 we are moving from a course to a smooth image. The variograms on the left provide us

with a very distinct understanding of these images. For a coarse image there is a large nugget, for a smooth image there is almost no nugget effect, right?

So, when you calculate a variogram or estimate a variogram given the shape of the variogram. For example, if you go back, if I have a variogram which looks like you know which sort of bends down to this value. It seems like the data are going to be quite coarse from the example that we were looking at previously right? And if you have a situation where the value will be small the nugget effect is small you are perhaps looking at a very smooth image.

So, the visual and graphical understanding that the variogram encapsulates is very clear from this slide. These data are coming from Perch and Dosh 20 14, but they are very informative in the sense of what to expect of a visual image if you look at the variogram and vice versa ok alright.

So, this image now says that we are going to move from an experimental variogram a calculation-based variogram to a variogram model. Now, a model unlike these mean scatter plots of you know square differences is a smoother representation of the variogram itself. Why should we go from an experimental to a modeled version of a variogram, well?

We should do that because looking for a given h value, we have infinitely many directions that we can pick if we were to really get serious about, the literal about. You know the experiment how are we defining our lags right? Even if I fix this h, I can just keep on changing it by delta keep on doing it, keep on doing it is probably going to be countably many many times before I can even get all the unique lagged pairs of data for a given value of h.

Imagine doing it for different or different values of h or distinct values of the lag vector h that is the distance h right? That is an ominous exercise we do not want to be doing that, it is like you know spatial prediction we cannot be sampling everything. It is because, if you were to sample a population then you know what is the role of statistics. What we are doing is we are getting some representation of the variogram at let us say the north-south, east-west, southeast, northwest, and then you know northeast, and southwest direction.

And that is about it we are going to then use it to generalize what is the variogram value at that given h no matter what direction you move into right? So, to be able to do that we need

to move to a model version of a variogram and that is the next step that we will start to look at in the next lecture. We will do this, we will study modeling variogram models in the next lecture. So, that is about it for today's, for this lecture, lecture 13, I hope this was enjoyable and knowledgeable for you.

Thank you very much for your attention see you next time.