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

Lecture - 13A
The experimental variogram

Hello everyone. Welcome back to lecture 13 of Spatial Statistics and Spatial Econometrics. In this lecture, we will study the tools that are needed to apply the theory of a variogram on a given sample data set. When I say that we will apply the theory or mobilize the theory of a variogram, I am including the theory of spatial stationarity that is when and how can we decide whether or not a given spatial domain is stationary right?

After that, we also include the idea or the concept of spatial contiguity. What is spatial contiguity and how do we measure it? What is the utility of spatial contiguity? Remember, when data are spatially contiguous that is to say that values at locations, which are spatially proximate or located closer to each other tend to be more similar usually than values at locations that are located further apart right?

Now, this phenomenon which is known as spatial contiguity is useful in spatial prediction. That is to say that we cannot really measure the values of any random variable. For example, air pollution right; that is the concentration of air pollutants in space or the crime rate in a city, groundwater levels in a region, we cannot possibly we going out and sample every single location in a domain of interest.

The domain of interest could be the world, in a city. It could be an administrative boundary like a district, a taluk, a village; you know a state, a country and so on and so forth. So, there are going to be many unsampled locations in space right? For these unsampled locations, we need to be able to predict the values that remain unsampled right? So, these are for these unsampled locations just because they are not sampled does not mean that there is no pollution well, there is pollution.

And if you talk about environmental justice, well people who are located near to nearer to areas, which were not sampled have an equal right to know about the quality of the air they are breathing than those areas, then those other households or people from whom the air

pollution monitoring station was quite nearby; that is they are nearer to the monitoring location.

So, spatial prediction is a fundamentally social scientific exercise right; in that form, that perspective which I just pronounced. So, spatial contiguity is a concept that sort of uses stationarity of a given domain and then, moves forward to then you know sort of provide us predictions at unsampled locations. And then finally, the variogram is a device that comes from the idea of intrinsic stationarity.

Another device that provides a measure for spatial contiguity apart from the variogram is the covariogram; which directly relates to the idea of second-order stationarity right? So, what we saw in the previous lecture is, first we were able to define the variogram and the covariogram in a statistically theoretical sense right? And you know we were also sort of able to say that the variogram is a preferred device over the covariogram because it is more general right?

It holds in its definition, holds for far more general settings than the covariogram. And that goes back to the theory theoretical understanding that you know space second-order stationary is strictly contained within the idea of intrinsic stationarity right? So, the variogram is preferred; it is a theoretically more interesting more general device for providing a measure of spatial contiguity in data.

And once we have this measure then we can take it forward and apply it to spatial prediction. And also spatial regression, as we will see going forward, but the question is how do we actually calculate or estimate a variogram device given you know a spatial sample data set right?

So, the basic quest for lecture 13 is, how do we calculate and estimate the variogram right for a given sample data set? Ok, now when I say calculate calculating a variogram, I am referring to something that we will see today called an experimental variogram ok.

When I say estimate, I will be talking about modeled variograms ok. And variogram itself is a device or measure of spatial contiguity ok. So, this makes it pretty clear the scope of lecture 13. So, let us move forward. What you have on the screen here is a textbook representation of a variogram, a covariogram, and a correlogram ok.

First of all, what is a variogram? Let us just do a little recall or little recap of the variogram right? So, a variogram is defined as 2 gamma h, where h is a lag vector and it is equal to the expectation of a random variable realization at location s minus the random variable realization at the location you know at lag vector separated by the first location s ok.

If we take this difference, we square it and we take its expectation. This is the theoretical definition of a variogram right? Now, in this theoretical definition, we always have understood this definition with this figure. So, I am going to just quickly draw this figure here, you have location S_1, which I am defining as s, and location S_2 which I am defining as s plus h right? The vector from S_1 to S_2 is called the spatial lag vector, spatial lag vector h right.

Like I have said earlier many many times that h encapsulates both distances, that is how far apart S_1 and S_2 are from each other and also, the direction from S_1 to S_2 right? In a spatial data set, both distance and direction are variable right. We do not have a convenient situation like a time series where the data are unidirectional and each hop from time period a to nearest time period b is equidistant right?

Or the distance between two time periods is exactly the same no matter what time, scale, or location you are talking about ok. So, having understood that the question is, how do we bring this definition to a data set? Of course, if 2 gamma h is a variogram, we also saw that gamma h would be a sem-variogram. In the textbook definition what happens is that if I start at location s.

And start moving from s to s plus h, which is separated by this lag vector h right. What happens is that this device provides me with a measure of spatial dependence between the two-time sample points. So, if h is very very small, if it is exactly equal to 0, here in this textbook variogram or textbook sem-variogram, what I see is that the variogram value is 0 or the sem-variogram value is 0 right?

And as we move further away that is as h increases the variogram value rises; that means, the smaller the variogram value the higher the spatial dependence. At location s, if I do not move at all and take another sample point at the same time, I am going to have a perfect correlation right? Because it is just the correlation of a value by itself right?

No matter how many times I sample this point right. I am going to do it, I am just creating a replica which is why the correlation is exactly equal to 1 right? That is when the variogram

value is very small, and as we keep moving forward the variogram value rises; that means, the spatial dependence falls and there comes a point when after which the variogram value will stop rising which is the point of no correlation.

The point of no spatial correlation; this point signifies that if I were to move further out from s to a distance, which is large enough let us say h prime from s, I will learn nothing from the value that is realized at location s for predicting the value at this location s plus h prime right? I learn nothing; this is no spatial correlation point, right? This is the point at which you know the height from the x-axis to this large scale with no correlation, you know, no spatial correlation point.

This signifies what is called the sill. The sill is nothing but large-scale variation and data. The distance h is signified here as range right. The distance R or when h equals R signifies the range after which there is no spatial dependence in data ok. So, this is a textbook variogram that we have seen earlier. This textbook variogram turns out to be a mirror image, a mirror image of a covariogram. What is a covariogram?

Well, we saw that as well, a theoretical covariogram is a covariance between Z of s, and Z of s plus h right? So, if there is high spatial dependence the covariogram value is pretty high. If after at the point when you know spatial dependence dies the covariance becomes 0. This is exactly what is happening with the mirror image as well right?

So, the variogram is a bit unusual from how we sort of study the dependence of two different random variables you know in traditional statistics. Statistics is nothing sort of more complicated than a mirror image of the covariance formulation or the covariogram formulation that we are aware of right?

When we come from the covariogram to the correlogram, which is nothing but C of h divided by C of 0; remember, this C of 0 nothing is a large-scale variation in data right? It is the large-scale variation in data, it basically means that this is the covariance or the correlation of a random covariance of a random variable by itself. So, covariance C of 0 is just covariance Z s by itself, right?

This is nothing but the variance of Z s and for a stationary domain, this is nothing but sigma square, which is exactly equal to the sill. This is interesting alright. So, the maximum value of a correlogram is 1. Well, that is the correlation between Z s by itself is 1 right? The

covariance is not 1, but the correlation is exactly 1 because it is just C 0 over C 0 when h is equal to 0.

If you want to sort of get a mathematical relationship between C h and 2 gamma h well, what you are looking at is the following. Let me use a different ink. So, it is clearer. So, we have 2 gamma h which is nothing but the variance of Z s minus Z s plus h is equal to the variance of Z s plus the variance of Z s plus h minus 2 covariances of Z s and Z s plus h. Now, by definition variance of Z s is C 0 look here right?

Plus again C 0 minus 2 C h. So, what we have is that gamma h is just C 0 minus C h. So, gamma h and C h are inversely proportional, that is why they simply mirror images of each other right? So, when at the point, when you know gamma h is 0 that is a point at the origin C h is just C 0. So, this height here is C 0, which is exactly the same as the height of the sill. So, this slide gives you a very sort of you know detailed understanding of what is a variogram, and what it really means.

How do we interpret a variogram, what is the intuition behind it? Using our understanding of the covariance from traditional statistics of correlation from traditional statistics, I highly encourage you to start with this you reproduce this slide at least a couple of more times by yourself. So, you get a very clear understanding of what a variogram is? a theoretical variogram is?

So, with that understanding, we will now move on to, taking a step forward to you know defining a variogram for a given sample data set. For doing that, let us recall this idea of local stationery from exploratory spatial data analysis. So, remember in case of ESDA right; in the case of ESDA, we plotted the values, the bivariate scatter plots of values realized at any given location with their neighbors one step forward and one step backward in different directions. You know in directions like towards North-South or East-West or you know both right?

And the idea of local stationarity is that these values should be similar to each other because they are located so close to each other. Now, this is also the idea of spatial contiguity, right? So, in this bivariate scatter plot, if I were to draw a 45-degree line that represents the area, where z of s is exactly equal to z s plus h. This is the line at which you know the correlation between z s.

And z s plus h will be exactly equal to 1 right? If all the scatter plot data sets were to lie on this line, then it will represent a condition when each data point exactly explains each other data point in its proximity. That is fantastic because then I can just sample any one data point and exact figure, what will be the next you know approximate data point to this observed data point, and then keep going one step forward and reconstruct the entire domain without even sampling more than one data point, right?

Well, that is not going to be you know that is an idealistic situation that is not how real data sets work right? The real data sets work like what we see for the coal mine data that we learned earlier. What happens with the real-world data set is that for any given value, any given real-world value, let us say we work with value 9. So, this is z of s equals 9; for this value when I go on to sort of you know drawing a vertical hash line.

It allows me to identify the values of z s plus h that correspond to this line ok. So, I have 1, 2, 3, 4, 5 you know approximate values to z s equals 9. Some of these values are very very close, and some of these are very close to z s. They represent very strong spatial dependence. Some of them are quite far apart, you know quite different right?

Very close meaning not location, but similar values right? And some of them are quite distinct or different relative to z of s equals 9 right? It could be in the positive direction or the negative direction irrespective of that the idea is there are high dissimilarities, right? So, on average, however, the data seem quite similar in their locality, but not exactly the same. Now, this idea of spatial dependence can then be brought forward to the variogram, right?

So, if I had only one z of s you know which is a very close value I expect the variogram value to be small, gamma h to be small right? And if I have a distinct value from z s equals 9 at the s plus h location right? So, then I expect the gamma h value to be high, that is reflecting a lower spatial dependence value ok.

Based on this understanding, let us try and create an experimental variogram. That is we will calculate a variogram value from the data.

To do that you know, first of all, this h scatter plot or this bivariate scatter plot that we have we have worked with; see that it reveals a correlation of data over a particular lag vector h, and we know that correlation comes from covariance, right? So, C of h is you know it should

be covariance. Sorry, about this typo it reveals covariance directly, and then correlation can be calculated as a function of covariance or correlogram, right?

So, it reveals covariance and we know from our first slide that covariance and correlation are mathematically, and graphically, mirror images to each other they are inversely proportional. That means, if I can get the covariance measure C h from the bivariate scatter plot, I should also be able to get the variance measure from this you know. Sorry, the variogram measure from this scatter plot is right.

Because we know they are simply linearly related right? So, the bivariate scatter plot can be a very good starting point. However, remember it is for a given value of h. So, it will provide me a 2 gamma or a gamma value, which is at a given lag vector h ok.

Now, given a spatial data set or a data sample, the variogram is written as the following. It is equal to it is given as 2 gamma h, which is by definition the variogram is equal to 1 over some value right.

The number of observations, which is denoted as modulus N in h. Summation, I goes from 1 to N h, summation of first difference squared values of realizations at location s and location s plus h. This value N, capital N h is a set. This is a set notation and it says it contains elements, i comma j in pairs. So, it contains pairs of elements such that they are separated by the lag vector h.

So, the set N of h contains all data pairs in our sample right? All data pairs, all unique data pairs in our sample are separated by lag or let us say spatial lag h, right? And the modulus value of h, this modulus of h is nothing but a notation for the count of unique elements in set N h. So, N h is something like a neighborhood data set, it is a neighborhood set, not a data set, it is a neighborhood set.

So, I am going to define it as a neighborhood set right? So, it sort of collects the pairs of neighbors that are separated by a lag h. The neighbor had set for the spatial lag h. Now, let us look at the data-based definition of 2 gamma h and the theoretical definition of 2 gamma h and try to figure out the correspondence between the 2. The theoretical definition was expectation z of s minus z of s minus h squared that is it, right? So, there is the expectation of the first you know sorry, the difference squared. So, it is a mean squared value of you know values at spatial lag h in my given data set.

The expectation operator has been replaced by 1 over N summation of these you know difference squared entities that is about it. So, we are looking at the sample representation of our variogram.

So, next after learning this, we are going to go through the process of actually calculating this variogram given the above formula right. I am going to write down the formula here again.

So, 2 gamma h is 1 over the number of unique elements in the neighborhood h neighborhood set for lag h summation i equals 1 to N of h. So, we sum through all the elements that are all the pairs of locations that are separated by spatial lag h summation of the difference of S_i from S_i plus h squared. So, I have a representation of a mean squared difference.

To do that to make our lives easy, what we do is, sort of the starting point you know called the tail, and the point that is separated by the tail is the starting point the initial point by a lag vector h is denoted as a head right. So, we are starting at S_1 and we are going to S_2, where S_2 is nothing but S_1 plus h.

Again h is encapsulating both the distance as well as the direction. For the given example, what you see here is that you have an h value let us say it equals 10 right? So, you take this tail and head representation and you collect the pair, where you have ZS_i and ZS plus 10. The colors on this regular lattice are basically representing you know data, instance data values, right?

So, they are just digital numbers that are now embedded as a color scheme right? So, you can note these values, you can take the first difference and then you can square it; that square enters you know right here. So, this is S plus 10 and this is S ok. Now, just like you sort of take this device, which goes from tail to head; what you do is, you take it around to every possible pair in this data set ok.

So, what you do is as a second step you are going to take it on to the next you know sill on the right and you are going to collect the pair k comma l. So, you can say this is ZS_k and this is ZS_k plus 10 right? The previous 1 was i. So, you will take this device and you will run it through every possible you know sample pairs that are separated by lag vector h.

Using these unique values, you will collect them in a set which will be called the set h right. So, we can collect all these pairs that are separated by you know the lag vector h, you can

imagine there will be very many such pairs. For these pairs, you will define the lag vector, the neighborhood set N h, and with these data and the set definition, we can evaluate the experimental variogram 2 gamma for a given value of h.

See that we are only looking at one h value in 1 direction. I can take the same h value and change the direction from east to west or west to east rather than north to south. That is to say that I could take the same vector and go downward from north to south and collect another you know sample of location pairs that are separated by the lag vector h.

Now, although h represents the same distance h equals 10, but in a different direction, and also nobody is actually stopping me from going diagonally in sense in the sense of direction. So, you can imagine that you will have many many values of h and many many values of you know in the sense of direction. Similarly, I could make my h values smaller instead of 10, I could be working with you know h equals 5.

So, then I will have many more sample pairs in these data. I could also you know equivalently work with a very large value of h, then you can imagine that my set N h with a large value of h. Let us say h equals 20 will have a lower number of entities than for N of h. So, for this regular lattice, I can claim that the count of elements or unique pairs that are separated by a lag of h h tilde is going to be smaller than the count with lag vector h.

If you know h tilde is greater than h and on the other hand, this count will become larger if the lag vector sort of I am going to look at the value, not the direction; if h tilde is less than h, right? I hope this is clear. This is just to make things clear, we are not going to use these things going forward ok.

So, this is for a regular lattice, but what if we have an irregular lattice in fact when we work with data we usually have to work with irregular lattices right? So, now, going forward in the next part of this lecture; we are going to study how to calculate a variogram if the data are an irregular lattice all right.

Thank you.