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

Lecture - 8A
Introduction to Spatial Autocorrelation

Hello everyone. Welcome back to this lecture series on Spatial Statistics and Spatial Econometrics. Today we are going to be covering lecture 8 which is a shift from what we have been doing till now. So, in the previous lecture or a couple of lectures, we have studied this idea of entropy and how we interpret the entropy of a random process to a spatially delineated random function space, right?

So, today we are going to sort of move slightly away from that you know that topic. And we are going to look at spatial autocorrelation and its consequence for the estimation of the sample mean. So, the first entity of discussion today is this idea of spatial autocorrelation. So, what is spatial autocorrelation? Spatial autocorrelation is a measure of dependence among random variables in space, ok.

So, you know till now we have talked about something like distance, we have talked about spatial dependence with examples, right? Now we are coming to a point where we are providing a measure for it, right? So, we are talking about a measure and this measure will likely depend on or will be a function of the distance you know between random variables in space.

Of course, spatial autocorrelation is what binds random variables in space and ultimately lends them to become a random function, right? We said that a random function is nothing but a collection of random variables which are jointly distributed. Now spatial autocorrelation provides a specific mechanism for jointedness in the distribution of random variables in space, right?

So, to make our you know understanding slightly more clear, we can sort of take a small example, right? We can say let us say we have random variable realizations for 3 locations in space. 3 locations in space are given as S_1, S_2, and S_3, ok. Now, the random variable that is realized let us say it is given by Z.

So, $Z(S_1)$ is the random variable realization at location S_1, $Z(S_2)$ is the realization at location S_2, and $Z(S_3)$ is a similar realization at S_3. We can also define these as simple indices Z_1, Z_2, and Z_3 where the index 1, 2, and 3 are delineating the location differentiation between these random variable realizations, right?

Now the point is that when we talk about spatial autocorrelation the idea is that first of all correlation between Z_1 and Z_2 correlation between Z_1, and Z_3, and a correlation between Z_2 and Z_3. These are the unique pairs in which theta occurs as in our example.

So, these correlation metrics are individually non-zero. So, there is a likelihood that either all of them are non-zero or at least 1 of these pairs exhibits a non-zero correlation. Now, this correlation is often you know a function of the distance between individual pairs, right?

So, if they are a function of distance that is why you know they are using an entity that is defined over space between these variables. And hence these are measures of spatial autocorrelation. Why autocorrelation? Because it is the same you know entity Z let us say its groundwater level.

So, if the groundwater level at location S_1, location S_2, at location S_3 are all groundwater levels, they are correlated with each other. So, it is called autocorrelation. So, its Z groundwater level at location S_1 is correlated with itself at the location at S_2 and the location at S_3, right? So, that is the point you know of spatial autocorrelation, right?

So, this is what provides you know a definition of spatial autocorrelation. So, it is a correlation in the traditional sense, but then as a function of distance in space, right? However, we define it can be Manhattan distance, it could be Euclidean distance, great arc distance, or even more sophisticated distance metrics that we discussed in one of the earlier lectures, right ok.

So, the next point is ok. So, let us say we are given these 3 data points. So, we are given these 3 beta points, right and we want to understand what is the mean, what is the mean, or the sample mean of you know these data realizations. So, the sample mean is straightforward you know all of you understand what I mean by that.

I am taking Z_1, Z_2, and Z_3 and I am summing them with equal weights of one-third or I am just summing them into by 3. Or you know I can simply say it is one-third times Z_1 plus

one-third times Z_2 plus one-third times Z_3, right? So, we are applying these equal weights because you know any of these values could occur you know when I am talking about the best guess of what the mean value would be, right?

So, any of these values are equally likely to occur in space. So, they are all realizations of you know they are all realizations of a random process. So, we say that the best guess offered the mean of this you know random process would be given the sample data set I have is simply add them with equal weights of one-third each, right? So, Z_1 plus Z_2 plus Z_3 by 3, ok.

This is what we call a Z bar and if we go and revisit the title of today's lecture, it says spatial autocorrelation and its consequence for estimation of mean. So, you know our aim going forward in this lecture is to evaluate the role of spatial autocorrelation in on Z bar. Spatial autocorrelation is nothing but a non-zero correlation among random variables in space, right?

So, we want to see if you have this correlation to this dependence metric you know how it affects the mean, right? Where does it affect, I mean at the mean? Does it at all affect it or not, right and. So, what we will do is to proceed we will first consider the case we are where they are independent, right? So, we will first consider the let us say consider the iid case.

What is the iid case? Which iid means independently and identically distributed or vice versa, ok. You can say identically and independently distributed random variables distributed, ok. So, independently and identically distributed random variables or vice versa identically and independently distributed variables, ok. So, let us begin our evaluation of that case, ok.

So, we say consider a sequence of iid random variables Z_1, Z_2, Z_3 go all the way till Z_n, right? Now here the indices 1, 2 till n represent location just like we saw in the previous you know slide, right? With the pointer that the random variable realizations at these locations are completely random, right? So, in the case of iid, you know with the condition that random variable realizations at these locations are completely random, ok.

What does it mean that means the value of value realized at location 1 that is Z_1 has got nothing to do with the realizations at locations 2 till n; that is Z_2, Z_3 till Z_n? Whatever those values or realizations are they have no bearing on what we realize on location 1 which is the value Z 1.

Similarly, what we realize on location n that is Z_n has had no bearing from what we had realized on location 1 that is Z_1 or Z_2 or Z_3 till Z n minus 1, right? So, that would mean that they are completely random and that there is no relationship or dependence spatially or otherwise, right?

So, we can say that is no relation or dependence especially. I am just going to focus on spatially speaking because that is what the scope of this course is, but otherwise what I am saying is they are completely random, right? So, where there is no other kind of you know grouping clustering going on in this data pattern.

So, mathematically we say we have a sequence Z_1, Z_2 till Z_n such that we say Z_i is distributed iid normal mean mu variance sigma square. So, now, there is a new piece of information that these are not only independently and identically distributed. But they are distributed according to the normal distribution with mean mu and variance sigma squared.

So, we know from here we know that mu hat is nothing but the sample estimate of the mean mu, right? So, mu hat is nothing but the sample estimate of mu is nothing but summation i equals 1 to n Z_i divided by n, ok. Which is also notated as denoted as Z bar, right? So, in the previous case, we had 3 you know realizations there we had Z_1 plus Z_2 plus Z_3 by 3 because their n was equal to 3, right?

Here we have a more general case, but the definition remains the same. We also know that sigma hat squared is the sample variance, you know representative of sigma squared which is the distribution population variance is equal to summation i equals one to n Z_i minus Z bar the whole square divided by n minus 1, ok. Now I want you to sort of ask yourselves or you know try and answer in the next 30 seconds where this minus 1 comes from in the denominator, ok.

So, I am asking why, minus 1 is the denominator. Ok. So, to be able to understand why we have minus 1 in the denominator well, we must recall the concept of degrees of freedom. What is the concept of degrees of freedom? Well, the degrees of freedom provide us an understanding or a measure of the net variation in data that we are using to define sigma hat squared, ok.

Now, when we write down when we define sigma hat squared we are indeed summing n entities, right? So, we are summing n entities, right, distinct n entities which are distinct due

to unique, ok unique Z_i's, right? Now the point is, but we are also using Z bar you know, but ok. So, we have unique Z_i.

So, we have Z_1 which is different from Z_2 which is different from Z_3 which is different from Z_4 through Z_n. Note that there can be similar values as far as realization it is possible that you know when they are realized as random entities. So, happens randomly that Z_1 is exactly equal to Z_5.

But they are still distinct entities that are drawn at distinct locations they are drawn independently they are drawn identically. So, they are distinct entities, right? So, they are not supposed to be the same if they appear to be randomly the same numbers, right? So, the sequence can have the same values at different locations, but they are still distinct entities by the virtue of location.

And the fact that they have been independently drawn you know at that location. So, we are summing n entities in the numerator which are distinct due to these unique Z_i's, right indexed by location i. But we also use Z bar in the definition and if you realize Z bar is nothing but a function of n Z_i's, right? So, here is the Z bar and it is a function of you know a Z_i's.

So, if the Z bar is a function of you know these Z_i's what it means is that you can give me any n minus 1 Z_i's and I will back out the nth Z_i, right? Because I know the Z bar in the definition of sigma hat squared, right? You can give me Z_1 till Z_n minus 1 and Z bar and I will back out Z_n. Or you can give me Z_1 to Zn minus 2 and Z_n and not give me Z_n minus 1 and give me Z bar, I will back out Z_n minus to the missing value.

You can drop out any 1 you know Z_i in the sequence of Z_1 to Z_n. And allow me to use the Z bar I can back out the remaining you know Z_i. So, what happens is due to this way that we have defined you know sigma hat squared where we use all i Z_i's, distinct Z_i's and we use Z bar. In the net what we are saying is we are only exploiting the variation of you know n minus 1 Z_i's, right?

We can live with just n minus 1 Z_i's and Z bar that is all the information that I need, right? So, the net variation that is being derived from the given sample to calculate the sigma hat squared just comes from n minus 1 Z_i values, right? So, I am going to just write here that one degree of freedom was lost due to the use of the Z bar in the definition of sigma hat squared, right?

This implies we are using the net variation in sample equivalent to n minus 1 distinct entity, ok. So, we are just on the net and variation is coming from n minus 1 entity, right? The net variation is coming from n minus 1 entity this is important. And you must go back and re-take relook at the idea of degrees of freedom, right?

But, it is really important to realize that you know there is an n minus 1 and y does this minus 1 appear, this concept will appear again and again in any statistics course, not just this one. So, this knowledge is kind of general, ok. So, the next important issue is the variance of the Z bar. So, now why do we have a variance of Z bar? Well, what is Z bar?

So, note that the Z bar is nothing but a best guess of the mean mu from sample information from what we know in the sample, right? So, Z bar itself is the best guess, right? It is only the best guess of what the population means mu is of mu given sample information, right? So, when I am given the sample all I know is that you know we have entities Z_1 till Z_n.

What would be what would the sample mean be? Well, it would be it could be equally likely that mu will appear to be any one of these entities, right? So, what do we do? We give each entity an equal weight. So, we equally weigh each Z_i that is with the weight 1 over n and then we simply sum them, right? So, we weigh each Z_i we get Z_i by n and we simply sum them and we get our Z bar.

That is our best guess that is our expectation of what mu will be, right given sample information. And that Z bar is also a random variable, ok. Why? Well, it is composed of random variable Z_i till Z_1 till Z_n, right? So, if you have a Z bar which is just a linear function of n random variables then of course, it is a random variable. So, the Z bar is also a random variable because it is a sum of n random variables, right?

So, it must also have a variance just like random variable Z_i have a variance therefore, the Z bar should also have a variable. So, therefore, the Z bar must also have a variance. So, till now we have just resolved that the Z bar will have a variance.

Now, we will go next to the next step and ask what will that what is that variance you know values. So, we will now next you know what we will do is we will derive this variance of Z bar.

So, the variance of the Z bar is nothing but the variance of 1 over n summation i equals one to n Z_i. Now, i will 1 over n it is a constant it will come out of the variance operator as 1 by n squared and I have a variance of Z_1 plus Z_2 plus Z_3 plus dot dot dot plus Z_n, ok. Now in evaluating the variance of the sum of n random variables, you know let us recall that the variance of 2 random variables x, the sum of random variables x and y is the variance of x plus y.

Is nothing but the variance of x plus the variance of y plus 2 covariances of x and y, ok. So, this is the covariance this is the variance of a sum of 2 random variables. So, let us extend that to the variance of n random variable. So, it will be 1 over n square times the variance of Z_1 plus the variance of Z_2 plus the variance of Z_3 plus dot dot dot variance of Zn plus 2 covariances of each pair Z_1 and Z_2 plus 2 covariance of Z_1 and Z_3 plus.

And you will keep going and get all the distinct pairs with you know to get a covariance twice the covariance. And finally, you will likely have Z_n minus 1 and Z_n, right? Now, interestingly all of these covariance terms are equal to 0 individually, and obviously, when summed together they are all 0. Because, because why? Well because Z_i's are independently distributed that is iid one of the is.

Says that Z_i's are independently distributed. So, what does independently distributed imply? Well independently distributed or this term independently the consequence is that the covariance of any pair Z_i and Z_j will be 0 if i is not equal to j, right? So, that is great. So, our problem is much more simplified.

So, the variance of Z bar is 1 over n squared variance of Z_1 plus the variance of Z_2 plus the variance of Z_3 plus dot dot dot plus the variance of Z_n, right? Which can be now written even more simply as summation i equals 1 to n variance of Z_i. Why can I write why can I sort of use this index? Because you know I am just summing the look the variance of you know random variable realizations at the location is.

And this can be further simplified as 1 over n squared summation i equals 1 to n sigma squared. Why could I do that? Because they are also all identically distributed, right? So, this is because you have identically distributed Z_i's with normal mean mu and variance sigma squared. So, all these you know Z_i's have the same variance sigma squared, right?

So, if I am summing sigma squared n times all I am going to have is n sigma squared over n squared. And finally, the variance of the Z bar is going to be sigma squared by n, ok. So, the variance of let us write it again.

So, the variance of the Z bar we have figured out is nothing but sigma squared by n, ok. So, now, the fact is that the sample estimate. We know the sample estimate of sigma squared is equal to sigma hat squared. So, the sample estimate of the variance of the Z bar from information that we have from the sample is given as sigma hat squared divided by n.

Where sigma hat squared was defined previously where we sort of had a little discussion on the degrees of freedom idea as well, ok. So, now, having sort of you know understood that we can write the standard deviation of you know we can write standard deviation of Z bar. Similarly, it will be sigma over root n and you know this will imply this and the variance of Z bar will imply standard deviation of Z bar to be given as sigma hat by root n, ok.

So, with that understanding now we will move to the next step which is called statistical variance. The statistical sorry, the statistical inference not statistical variance I am sorry about that. So, we have the next topic being statistical inference. So, the idea is that you know we are trying to get at the sample you know estimates of mu and sample estimate of you know sigma squared. Now the sample estimate of mu which is the Z bar itself is a random variable.

So, it is going to be it is our best guess, but there can be some errors, right there can be some errors. The idea is that I have data that is all distributed according to the same distribution which is normal with mean mu and you know variance sigma squared, ok. Now you have data sequenced such that you let us say you have Z_{10} here you had Z_1 here Z_3 you know Z_4 you had Z_5 here you had Z_{51} here and Z_6 and Z you know Z_8 here and so on.

So, have the sequence of values which you know play somewhere on the real number line. And the idea is that you have a normal distribution with mean mu and you know variance sigma squared. So, the best guess of mu that we have is the mean of these values which is likely to be somewhere here, which is the value Z bar. Now, this Z bar as we said is the best guess of all the different realizations of Z_i's that I have in my sample, right?

The idea of the Z bar is that it can take any value with an equal weight between the in this set of Z_1 to Z_m, right? So, the idea is that the Z bar itself is a representation of what we expect the mean to be, right and there could be some error. So, we are aware you know of a strategy to

articulate this error or possibility of error in my best guess of mu and the way I write it is that you know that we can say that mean value lies in the following interval with 95 percent confidence, ok.

What is the idea of 95 percent confidence? The idea is that look I mean I am going to give you a best guess based on what my sample information is, but it comes with some error. So, I am not going to restrict myself to just Z bar. I am going to give you an interval within which you can understand the meaning of life, right?

So, the Z bar value itself for a different random sample with the same distribution you know understanding might have slightly different values, right? Depending on what sample information appears, we can get a sense of what that value will be with 95 percent confidence.

And we know you know if you have studied basic statistics you would recognize that you will you can write this value as Z bar minus 1.96 Z hat by root n and Z bar plus 1.96 sigma hat by root n, ok. So, what I am saying is that, ok you know this is the standard deviation of the Z bar.

So, what I do is I go from Z bar 1.96 sigma hat by root n on the left. I go the same distance on the right and I say that the Z bar is going to be lying in this interval with 95 percent probability, ok. So, now, I do not provide you a point estimate I provide you an interval estimate, ok.

So, values that are outside of this bound. So, let us say you know you had valued at 10 and you ask ok do you think Z bar could be equal to Z_{10} you would say no it's outside the 95 percent you know confidence interval. So, I guess that the Z bar will not be equal to Z_n. So, that is the statistical inference we are using the idea of errors to get at statistical inference.

So, this 1.96 times sigma hat by root n is called an error in our best guess, ok. And where does 1.96 come from? Well, you should be there I will say recall the t statistic, ok. The students, t-test, the t statistic and I encourage you to go back and read about it if you want to understand where the where does 1.96 come from.

If you go to a t table on the back of any statistics book and you look for you know the critical value of t for 95 percent confidence or 5 percent error, that is a 2-sided 5 percent error. You

will you know see that for very large degrees of freedom, the critical value turns out to be 1.96. So, you can locate this on a t table go and locate it go and read about t statistic it will become clearer.

Another interpretation of this idea of this range is that 95 percent. So, alternative interpretation, and then we will move to the next step, ok. So, what is the alternative interpretation? The alternative interpretation is that 90 percent of the data in sequence Z_1, Z_2, Z_3 till Z_n lies in the interval, lies in the interval the same interval Z bar minus 1.96 sigma hat by n root n and Z bar plus 1.96 sigma hat by root n, ok.

So, just like in the previous example on the previous I mean the example on the previous page you saw that Z_{10}, and Z_1 are outside of the confidence bound. That is Z bar minus 1.96 root sigma hat by root n and Z bar plus 1.96 times sigma hat by root n. So, there are 2 values there will be about 5 percent data in the sequence that we have which will be outside these bounds those are data points we will say that is probably not going to be the values of Z bar, right?

So, the Z bar will not take those values in all likelihood even if you gave me a different random sample which was iid n with the same mu and same sigma squared, ok. So, this is statistical inference and now having understood this idea, we will next what we will do is the next step that we will take is that, we will introduce spatial autocorrelation to this to our example.

That is an example sequence Z_1, Z_2 till Z_n and evaluate. And evaluate the impacts on mean and statistical inference based on the mean on it on the mean itself, right? So, let us do that in the next you know in the next module.

Thank you.