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

Lecture - 12A
Spatial Dependence

Hello everyone. Welcome to the 12th lecture on Spatial Statistics and Spatial Econometrics. Today, we are going to sort of steer in a direction that will provide a more definitive measure of spatial dependence. We will obviously, mobilize the concept of spatial stationarity that we covered in the previous lecture.

If you think about it we have covered quite a bit of ground till now, right? So, if we sort of you know before moving forward to a quite substantial you know substantially different direction, let us do a short recap to appreciate the ground that we have covered till now and how sort of it binds together and provides a pathway for the next you know chapter in this series of lectures.

So, if you were to do a short sort of recap short and quick recap, first you know we started this course with an introduction to modern data sets, modern spatial data sets, and their applications, right? So, we sort of you know sort of looked at who cares about spatial analysis, who uses which sector and the in the industry, has extensive use of these you know of this toolset and where does it fit in academic research and so on and so forth.

We also saw that you know now we have quite many freely available, ready-to-use spatially delineated data sets, right? After that, we developed what we called a general spatial model. A general spatial model basically provided us with mathematical devices to summarize the data that we see pictorially or visually, right?

So, we are very used to looking at spatial data as an image, right, as a picture that you can click on a camera of your phone or you know very sophisticated cameras you know that may be handheld, could be mounted on a drone, or even you know attached to you know a satellite as a sensor, right. All of these data sets can be sort of you know summarized using a general mathematical model, right?

And when we did that we looked at things like distance metrics, right, different distance metrics that help us summarize you know the spatial patterns in these data sets. And we also

looked at what we called the data structures, right? So, data sets come in various you know structures, and we also summarize some of them as vector data, and raster data, within those there are several other forms you know we have these definitions called geostatistical data, lattice data, right so, and so on.

So, vector and raster are formats to store data whereas, you know geo statistical or lattice data or point data, polygon data so on and so forth are you know are ways to manipulate or analyze those data. So, the structures have different forms and meanings, and uses, right? After that we changed our cares a little bit, we looked at this concept of spatial entropy, right?

We took entropy as a measure of variation in space, right, general variation in space. And then we use this pattern to provide a model for a monocentric city, alright. And after we covered spatial entropy which is a measure of variation over space, we move one step further and studied something like spatial autocorrelation, ok.

So, from a general measure of variation in space, we went on to talk about a measure of dependence in space, right? And we looked at spatial order correlation and its consequences for statistical inference. We did this in 1-dimensional with a one-dimensional you know sequence of data; we also did this for a 2-dimensional sequence of data.

Then we went on to you know also study you know I am just going to say 4 continued, we also looked at what we say called Monte Carlo simulations, right? We said look I mean you have all these mathematical analytical results, which can be quite complex at times they involve you know applying algebra, calculus, and various other concepts. A numerical substitute or an alternative for these analytical results that we have, that we study in statistics our simulations, right and specifically Monte Carlo simulations.

After that, we covered what is called exploratory spatial data analysis, right. We said that whenever we get a data set, our first job is to summarize that data set. What if we get a data set that is spatially delineated? So, we covered methods and tools that are specifically designed for such data sets.

And just in the previous you know couple of lectures, we have looked at the concept of spatial stationarity, ok. Spatial stationarity is crucial. It is a decision. It is not a hypothesis, it is a crucial decision. It validates any statistic in space, right? Even mean, median, and

outliers, whatever we do we must work in a stationary domain to conduct any kind of spatial analysis, ok.

So, now, we have come to a point where we can indeed understand a stationary domain and we can conduct spatial analysis.

We will now go on to you know formally introduce the idea of spatial contiguity or spatial continuity and provide a measure for this or alternative measures for spatial contiguity over stationary domains.

So, going forward whenever we define these concepts including the means and variances and you know measures of summary statistics or spatial contiguity or spatial order correlation, we are going to be assuming that we are working with a stationary domain. And wherever we have, you know, a discussion is due to a sort of make sure that we are indeed working with the stationary domain, we will you know provide that discussion, ok.

So, let us move, right let us sort of you know move forward. So, spatial contiguity you know could also be called spatial continuity. Ultimately, we are thinking about spatial dependence, ok. We are, right, we are thinking formally about spatial dependence. So, let us figure let us start with you know figure in front, and let us talk about the panel on the right. So, let us look at the right panel first.

So, the right panel provides me data that are visually delineated with 5 different colors, ok. These colors are you know green, gray, orange, black, and blue. Say these colors represent different levels of intensities at which the data are recorded. Often these intensities are also called digital numbers, right? So, we assign a number, a unique number to each color. So, that you know each color sort of is providing a categorical level at which the data are you know observed.

Quite clearly on the right-hand side what is happening is that you have data distributed in columns where the data moves first at level 1, then level 2, and level 3, and then level 4, it's going to and then level 5, alright. Now, what is the utility of having data distributed over space on the right-hand side the way it is that there is some kind of a spatial structure going on?

And the same spatial structure is that when we begin from left to right, we begin from left and go on to the right direction, what happens is that the data remain at the same intensity for a certain distance before it crosses onto a boundary after, which it sort of changes its intensity to something else, ok.

That means, that there is spatial dependence going on in a left to right, such that it sort of changes drastically discretely after a boundary. Whereas, from you know top to bottom, if I start at any given value and I straight go south, I am going to be on the same value. This tells me that if I have an observation at any point on level 1, if I want to predict an unknown value you know south of the observed value, right of I at south you know I can accurately predict it to be belonging to category 1 or level 1, right.

So, spatial dependence as categorized in the figure on the right-hand side, the right panel, right, is providing me an opportunity to utilize the fact that the data are spatially dependent in a given structure to then predict unknown values from observed or sampled values, right? We talked about the fact that we cannot realistically sample every location in space, right? It is very expensive, right? And often not even possible because you know all areas in space on the land for example are not accessible, right? So, prediction is key, right?

And if you have some spatial structure, some spatial dependence structure, right? which provides you know a kind of a correlation structure over space, right a definitive correlation structure, then we are looking at a scenario where we can use this information to predict values of a random variable at unknown locations, right.

So, if coal were to exist as the right panel, right it makes my life very very easy, right? I can make samples you know in a direction, right, from left to right. And I can find these boundary points and once I find these boundary points I am kind of done, right? Once I have my boundary points figured out over space, I do not need to sample anything else in these columns alright, right because these columns are perfectly correlated from north to south.

So, all I need is an estimate of where the boundary points are, right? If life were so convenient in the real world you know prediction would become very very accurate and very very easy, right? But that is not going to be the case that is not to be the case. So easy; you know it does not appear to be so easy in the real world.

If you look at the left panel that is a case where the data seem to be distributed or the categories or levels seem to be distributed more randomly in space, right? That is to say that if I were to take samples at some locations in space and use those to then you know try and predict unknown values, it is going to be not so easy, right?

So, some of the values are missing, and some of the values are known, right? So, the circles are missing data, and my objective as an analyst is to be predicted, to predict these data and crosses are where I have my sample observations. So, I can merely use the data that are observed at these you know the locations where I am sampling them, and I want to predict you know what is happening at unknown locations, right? It is quite possible that some of the locations might not have any known samples.

And then if there are so many boundaries to be predicted in so many different directions, it is going to make my life very very hard to figure out, you know, what is the structure spatial dependence structure do we have any? Of course, there is a lot of very strong spatial you know dependence within each block or each polygon so to say. But to predict where the boundaries of these blocks will sort of you know end or we are going into a different regime, I am going to have to do more intense, more costly sampling.

And I really do not know if this structure on the left panel exists beneath the ground I really do not know how many samples to take where, it is a very difficult abstract problem. So, the panel on the left represents a scenario where there is no or little spatial contiguity or spatial continuity, right?

That is there is little or no, you know very less correlation in data on the left-hand side, you know across samples in space and you know you are going to, you are looking at somehow random values being observed you know at different locations in space, right? If I were to go in and sample locations, sparsely I will probably merely get data you know in all different you know regimes.

They are going to be very different values. I am going to be confused when I look at them I am not going to be able to find a structure, and you know honestly it is going to be very very hard to predict what the unknown values are with a sparse sampling exercise, ok, right.

Everything that is non-sampled is actually unknown, right? So, the circles are everywhere we do not have crosses. So, you can imagine the kind of intense exercise it would be to predict to

conduct spatial prediction when the real-world natural structure is given as a left-hand panel versus the right-hand panel, right?

And it is hard to learn from sample observation. So, spatial dependence, spatial contiguity although when we look at it you know it as a concept to learn separately and as something to estimate, it seems like you know it is more work.

But on the other hand, it provides us a utility on the side of prediction. That is if I know a few values, if I know a few values at the boundary you know I can know everything else without doing much, right? So, some pain in estimating you know spatial dependence structures, but a lot of value in terms of spatial prediction, ok.

Now, let us sort of you know define spatial contiguity. It is a summary of spatial pattern as a univariate statistic, just like mean-variance, but beyond mean and variance. It is typically a measure of the correlation of values over space. Values are observations, right, and I say values I mean, mean data values, right?

So, observations in space and it is a measure of how similar or homogeneous these values are over space. Spatial contiguity provides information that will be useful for prediction in space, right? So, it is the sort of the utility, the utility of going through this process of estimating spatial dependence of you know learning about spatial dependence in data, ok, alright.

Let us move forward. Let us look at this much more sort of nicer example by Professor Michael Pyerz. Now, on the left-hand side you see a random process, right? I mean you have these little little little cells over which the data is distributed.

And if you are standing on any cell it's hard to predict you know what is going on in its neighborhood because that neighborhood is so heterogeneous, it is so different no matter where you stand on this left panel, right. So, you have you know a very high spatial heterogeneity going on, even local spatial heterogeneity going on the left panel.

On the right panel, however, we see a very non-random spatial process, right. On the right panel what happens is if I stand at any location, I have some level of confidence that you know around me things are going to look very very similar, ok.

Of course, there are complications because there are these boundaries over which the process is changed, but you can imagine that you know some a not so intense sampling exercises can

provide us with where the boundaries are. And once we know the boundaries we will know you know what to expect of the values in the neighborhood, right?

So, again an example where the left panel provides us with no spatial dependence. So, you know if you know spatial statistics probably, you do not you know get too further in the sense of estimating the structure of spatial dependence. But then you also do not have any information for prediction.

On the right-hand side, you can have a situation, which is much closer to spatial processes, right? If you think about long-term geological processes, the way you know land structures, elevations, dams; not dam-sorry river structures, you know river pathways come about you know those are you know glacially glacial processes, right?

So, they have been happening over time, they are very slow and they are also highly spatially contiguous. You do not have plains and mountains moving you know mountainous regions you know fluctuating over space. You have a mountainous belt and then you have a large plain structure just like if you think about the geography of India, right?

So, learning about spatial statistics indeed sort of provides us with an opportunity to model such structures, but then also to conduct prediction which has a lot of economic and social value, ok. Now, here you know it is an example where we can consider a decision to harvest groundwater from an aquifer having different permeability measures, right? So, an aquifer is something that we saw last time it is a rock structure beneath the ground and it could be permeable or porous or non-permeable or non-porous, right?

The colors on the right-hand side you can think of them as providing a gradient from permeability to non-permeability, right? So, reds could be highly non-porous rocks, just like you know tubs where water does not move around laterally, and blues, I mean moving from red to yellows to greens to blues we are moving towards more porous structures.

Now, if we want to understand where should we draw groundwater from or where the depletion could be a lot bigger problem, well such an understanding of the structure beneath the surface will be very very useful. And in the sense that you know if I draw water from such let us say location where I have sort of mark on your screen, you know there will be a larger spillover on this you know porous structures except for this little red rock in the middle you know if I were to draw a lot of water from this you know southwest well.

If I draw water from the central well, it is going to cause a lot of spillover in it you know the northwest and western directions, and very little in this non-porous direction which is the southeast direction, right? So, the spillovers in groundwater levels that are depletion or declining will probably happen more and more in the locality in not, in directions that are not the northeastern direction where you have these high non-porous rocks, right?

So, there is no, they are not going to allow water to flow in if you start drawing a lot of water from their central well, ok. So, such economic decisions are possible if we have spatial dependence in space or spatial contiguity in space.

What I present here is another very nice, I think a very nice you know way to sort of understand if you see data if you have a univariate data set where you have porosity measures and depth from the ground or you know from any location depth going in any direction over space. Then, if you plot these data and they look highly irregular, right they look highly irregular then that is a signal of low spatial correlation in the data.

Low spatial correlation means less spatial contiguity that is more like left panels of the two figures that we have seen and that means, lesser information for prediction. It is also clear the fact that you know if you were to sort of you know observe let us say these three observations that I have marked with the cross going down, it is going to be very hard to predict what is coming next, right?

Worse is if you were doing it for example, in the case of a very high spatial correlation structure which is the figure in the third, in the second quadrant, right? which is on the southeast of this image. Then, you know predicting you know will be easier if you had you know unknown samples you know relative to the figure on the northwest you know of this image, ok, alright.

So, the next step is now that we understand spatial contiguity, its utilities, and so on and so forth. The next step is measuring spatial contiguity.

And the good news is we have already seen the devices you know most of the devices that are involved in measuring spatial contiguity. First is the variogram, now when we studied intrinsic stationarity, we specified a variogram in the definition of intrinsic stationarity. So, the variogram is written as 2 gamma s_1 minus s_2, s_1, and s_2 are vectors, and s_1 minus s_2 means

the distance between these vectors, so distance metric and location. And they are equal to the variance of the first difference between values observed at this distance, right?

We also looked at the covariogram last time which is nothing but the covariance of the values observed at any two locations in space. We also have a correlogram which is nothing but a correlation analog with you know univariate statistics. When you move from covariance to correlation, we move we can imagine moving from covariogram to a correlogram, right?

So, we will see, we will go over these devices or these measures one by one, but I have some notes for you on the right-hand side of the screen, which I would like to go over before we go there. So, all these measures are defined over a domain of analysis, right? So, as analysts, we choose this domain. We have talked about this quite a bit in this lecture, in this lecture series, right?

So, this D must be a stationary domain, right for any and all of these statistics to be valid. If we are working with a non-stationary domain D, then we cannot define a variogram, right? So, the variogram remains undefined. C 0 is the variance of random variables at a given location. So, C 0 is just the covariance of Z s_1 and Z s_2, where s_1 is exactly equal to s_2. If s_1 and s_2 are exactly equal, the distance between them will come become 0 and this covariance will just be the variance, right? This is something we are already aware of.

The quantity of gamma is called the semivariogram. Variogram is a measure of dissimilarity over a distance; well we will look at it more carefully in a couple, of in a couple of minutes. And the covariogram is a measure of similarity over distance. So, what I am declaring right away is that variogram and covariogram are directly inversely related, ok.

This is also something that we have seen earlier, but now going forward we will now develop these things more formally and look at their statistical properties, ok.

Now, one thing that we see here is you know a lag vector h. Now, h is a vector as seen during discussions of stationarity again is the L 2 norm or the distance metric between two locations location vectors s_1 and s_2, right? h always and always will contain two pieces of information one is the direction, right where is the set, when we look at a pair of observations what direction are we moving when you are moving from s_1 to s_2 or s_2 to s_1 and vice versa, right? And the other is a distance, right?

And we here what I have done is, I have simply sort of taken the definitions of variogram, covariogram, and correlogram, and you know I have given you the definitions in terms of h rather than s_1 minus s_2 in the vector form, ok.

So, going forward, we are going to now sort of we are where we will discuss the variogram in detail.