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

Lecture - 15A
Non-stationary spatial domains

Hello everyone. Welcome back to another lecture on Spatial Statistics and Spatial Econometrics. This is the 15th lecture of this series for spatial statistics and spatial econometrics. And, today we will be moving forward from the radiogram, you know theoretical variogram and experimental variogram and a variogram model, where we had this you know we got to a point where we have a generic device to which if I provide a spatial lag, measure or spatial distance between any two points, that is h it provides me a measure of spatial contiguity between those two points right.

So, today we are going to move one step forward and look at an example, starting with an example where our basic aim is to predict unknown values in space right? So, we have been building up to this point to be able to predict, one of the objectives is to predict unknown or unsampled locations in space. And, then from there, we will go on to the next module which is spatial regression wherein we will take this leap from you know correlation to causation as well right?

So, let us move forward. So, I have this slide titled spatial dependence estimation and Prediction for a Non-stationary Domain. As I said, we are going to work with an example, but in this example, we are going to assume that we are starting with a non-stationary domain. So, you know we are working with a real-world problem, it is not going to be a textbook representation, it is going to have issues of its own.

So, you know as an analyst how do we go about tackling these issues systematically, eventually getting to a point where we are using a variogram model to be able to provide a spatial prediction? So, we will start with an example and then finally, towards you know after this lecture, after lecture 15, you will have learnt also the theory of spatial interpolation or basically spatial prediction.

So, we have to start with a non-stationary domain. I am going, to begin with, an example, of groundwater levels data in 1-dimension right? So, 1-dimension means it is the real number

line. So, basically, we are based, we are walking on a real number line, and at different locations, we are measuring the groundwater depth.

So, let us draw that representation here. So, we have a ground surface, on this ground surface I have locations S_1, S_2, S_3, S_4, S_5, and S_6. At each location, I have a realization that Z of S_1 equals 2. So, 2 meters in depth right? Z of S_2 equals 5, Z of S_3 equals 12, Z of S_4 equals 8, Z of S_5 equals 15, and Z of S_6 equals 20. Now, even before we do anything, you know we start working with these data; it is pretty clear that as I am moving along the real number line from location 1 to location 6, the depth of groundwater level is on average rising right?

There we sort of sense a rise in these values for this particular example right? And, this X axis, I am just going to say the X axis is my ground level right? This is my way of sort of you know this is where I go in dig a well and figure out the value of well depth and you know Z of S_1. So, Z by itself represents groundwater depth, just for completeness. So, we are going to draw these data on a graph.

So, we have let's say groundwater depth on the Y axis and on the X axis I have my locations. So, I have my S_1, I have my S_2, S_3, S_4, S_5, S_6 and S_7. So, I do not have a S_7 so, I am going to stop at S_6. And, let's say on the Y axis, I mark level 5; so, this 5 is groundwater depth. So, on the Y axis, I have ground water depth let's say it is in meters and X is also in meters.

So, as I am moving from S_1 to S_2, the distance between them will be given in meters, and let's say 5, 10, 15, 20, and 25. So, at S_1, I know that at S_1 the value observed value is 2. So, it is somewhere around here, at S_2 the value is 5. So, it is somewhere here. I am just going to mark them, you know slightly with some approximation. But, I will try my trial doing my best to my ability to mark them to scale ok. So, Z is 4, 5, and 6, which is 20. There we go ok.

So, as I saw when I looked at the numbers, I felt that as if when I am moving from S_1 till S through S_6; you know the groundwater depth is sort of rising. So, the groundwater situation is becoming worse and worse as we traverse from S_1 to S_6 right? When I plot this data, I am able to work out a trend between these values. So, if I were to try and plot this, I am going to have something of this.

So, this is the spatial trend in groundwater data and this is not this figure although is in space should not surprise you, imagine the groundwater data was some kind of a time series. Let's say it was GDP and S_1, S_2, S_3, S_4, S_5, and S_6 were time points. So, then you know I am

looking at a time trend. So, the trend is not something that is very new to us when as statisticians or econometricians right? But, to be able to look at a spatial trend and to plot it, that is perhaps you know quite new ok. So, let's see what this trend effect really does.

So, if I look at these values, I mean I should be able to also figure out what you know the G bar that is the mean value of groundwater depth for these data is 2 plus 5 plus 12 plus 8 plus 15 plus 20 divided by 6. I believe if you sum them up, it will be 52 by 6. So, it will be somewhere around 9 meters ok. So, this is something we are used to starting with some kind of summary of these data.

I have just pointed out the mean groundwater depth in the domain that I am studying. So, my domain of interest is S_1 through S_6, right? So, my domain of interest DE comprises of locations, well we can say locations you know S_1 through S_6 ok. Even though S_1 through S6 are discrete points, every point between them is the potential for observing groundwater levels. And, the locations where we do not get to measure or you know sample these data, well those are the locations where we should be able to predict these data.

So, this exercise aims to be able to say take a point between S_3 and S_4 and call it S_0. And, we want to you know our aim here is to estimate or predict the value of groundwater depth at location S_0 right. This is what we were after, this is a prediction exercise. S_0 again is a representative unknown or unsampled location. Every other location that remains unsampled is a candidate for prediction, right?

So, what do we do? Well, as step 1, you know what we will do is we will model spatial dependence in the data using the given sample right? So, when I say data, I mean the data that is given to us, the data set that has been given to us. Why do we do that? Well, if we are able to predict spatial dependence right, remember I am looking for a point S_0, where there is no observation. I want to know where would S_0 the value of groundwater level or depth of ground level, what will it be at S_0?

What has been happening before S_0 is that there is a rising trend from 2 till 12. So, it is a very quick rise, but as soon as I come to S_4, it seems like there is a drop. So, this is pretty confusing as a statistician. You know if they kept rising, I could have said, you know maybe these values, the value is between you know 12 and the next higher value right? But, what happens is that S_4, it comes down, GW comes down and at S_5 it again goes up drastically right?

So, I do not really know, it's really hard, and it's non-trivial to predict what should be groundwater depth at location S_0. Now, we want to model spatial dependence so, that we will be able to know what is the strength of dependence around S_3 and around S_4. If the spatial dependence is very strong around S_3 and not so strong around S_4, then S_0 will be nearer to S_3 in terms of its representation of groundwater depth.

If it's vice versa, it might be that you have you know a greater spillover coming from S4 than from S3 right? Clearly, S0 is closer to S3 and slightly further away from S4. So, that will also be accounted for us. So, there is closeness and there is also this spatial dependence, the strength of spatial dependence at those two locations; that is what is going to be you know revealed by the variogram.

So, we will use a variogram and more specifically a variogram model, right? Apart from that, I would say clearly the knowledge of this trend, the knowledge of this trend of spatial trend in the X direction, in the positive X direction will help or helps or aids prediction right; help in predicting groundwater level at S_0 right? However, this very trend will make our data non-stationary right?

So, however, the trend effect makes our data non-stationary. We will just look in a minute at how trend makes our data non-stationary. Now, having this trend effect, this kind of structure in the data which is coming in as trends, it could be a non-linear trend as well right; other than that you might have regimes right? It might be that S_1 through S_3 is data in regime a, and S_4 through S_5, S_6 you have a data set in regime b right?

Whichever, way this structure arises right to a continuous trend or a piecewise average or piecewise linear representation; this structure provides me some knowledge that I can use then predict the S_0 value right? But, on the other hand, this makes my data non-stationary which means all my definitions of G bar, and my definition of variogram are useless. So, basically, the variogram is not defined because the variogram requires intrinsic non-stationary to hold for us to be able to write that, you know second-moment property of spatial data ok.

So, there is a pro and there is a con of spatial non-stationary data. And, this sort of you know provides a trade-off due to data structures, structural breaks or patterns, and data. Just a remark so, that it sort of you know stays with us ok. So, with this now that we have an understanding of this trend, let us go ahead and try to sort of model this trend.

So, say we model the spatial trend and when I say that what I mean is that you have groundwater level at location S_i, I am writing that as beta 0 plus beta 1 i. So, i is the index for space. When I move from S_1 to S_2, this movement from 1 to 2 tells me that S_2 is located in the positive X direction from S_1 and similarly then S_3, S_4, and S_5. So, as the index i increases, I am moving along the X-axis.

So, this is a convenient specification that the data structure allows me and then I have a random error. So, this is called a linear trend regression. Again, those of you who have seen time series analysis must have been definitely aware of this. This is a simple linear regression model. So, we are introducing a regression model, we are modeling the variation in G of S_i which is a spatial variable.

On the left-hand side, we have a linear model of what I want to explain right? So, my dependent variable is groundwater depth at location S_i which has a systemic portion of the variation. So, systemic portion or component of groundwater evolution in space right? How converter levels evolve in space, you know is given by this systemic portion.

This exhibits, you know spatial trends in data right? The parameters beta 0 and beta 1 are called model parameters, right? So, beta 0 and beta 1 are model parameters. Beta 0 is an intercept that is a representation of the global mean, a representation of mean values. Beta 1 is the change, the marginal change on moving to one step from you know lower X value to a higher X value right?

So, it is a step change in groundwater depth upon moving from location i to location i plus 1 right? So, the change in G S_i between location i and location i plus 1 will be driven by this systemic portion eta 0, u_i on the other hand is called the random error. This is the variation in groundwater levels, that is not explained by the trend which is not surprising.

You see the trend you know is one way of learning about what is happening with groundwater levels in space. There are so many complex mechanisms or processes in place that deliver the groundwater level that we see. Well, there is a draft, you know there is the extraction of water at different locations. There is a recharge in different locations.

There are anthropogenic impacts, there is rainfall whether you know rainfall is the same in all locations S_1 through S_6 or you have some higher level rainfall regions and some lower level. All these factors that can explain groundwater level which is different from a mean and the

trend effect which we have included in our systemic portion, will all go into the random error you know u_i. It is called an error because this is the error of our model. The model is beta 0 plus beta 1 i, right?

And, this error is one where we assume that at on expectation, this error will be 0 ok. So, very quickly if you want to see, if you want to go back to the previous slide where we drew this data; well you know this trend line has an intercept. And, it has a slope, a slope is the step increase in moving from i to i plus 1. So, this is beta 1 representation right on your screen. And, you see that the predicted value at S_6 from this model will be beta 0 plus beta 1, you know times 6. It will be a value on the yellow line.

Now, this predicted value which is in black and the truth which was the actual value observed value which is in blue are different right? This difference represents the error from our model ok. This is the error in the regression, linear trend regression on the previous slide ok. So, let us move forward ok. So, now, I have my understanding of this model through spatial regression.

Why is it spatial regression? Because it is explicitly exploiting space to specify the random variable that is a groundwater depth in a given region or domain d right. Now, straight away you will see that if I were to write the expectation, you know the expectation of G of S_i, it will be the expectation of beta 0 which is a constant that is just beta 0. So, the expectation of a constant is the constant itself plus you know beta 1 which is again a constant and i is deterministic.

So, location is not random, i is when I say location is a 6, I mean it is exactly S_6, not like it is S_6 plus minus somewhere. There is no such measurement error or any kind of error in location i. So, my expectation you know of G S_i is just beta 0 plus beta 1, i because the expectation of u_i is 0. Similarly, if I were to go to a separate location S_j, here the expectation will be beta 0 plus beta 1_j.

This implies that the expectation of G of S_i will be not unequal or not equal to the expectation of G of S_j, for all i j pairs such that i is not equal to j. So that means, the expectation, the first moment that is the mean at different locations is different due to the trend effect, specifically due to the trend effect right? This is where the non-stationary is coming from right? This is the definition, the first condition of stationarity be it intrinsic or second-order stationarity is that the mean is the same at all locations.

Well, that is not the case here. So, that is why the trend leads to non-stationarity. So, let us move on from the first moment characterization that leads to the reason for non-stationarity to the second moment characterization which is the variance of G of S_i. Now, the purpose of a regression model is to ultimately be able to explain the variation of the dependent variable.

So, ultimately as an analyst when I am specifying a regression model, I am trying to explain the variation in groundwater levels right? So, if I go back, you know if I look at the raw data, there is this variation in groundwater levels that is happening. Of course, there is a trend effect right, that is why I include it in my model. But, ultimately I want to be able to do my best to explain this variation in groundwater levels which are going up and down through space right?

Now, so, second-moment characterization is about you know what is the total variation in the data. The total variation in the data will come from the second moment, that is the variance of G of S_i. Now, part of this variation will be explained by the trend right? The trend characterization is a systemic portion as we have specified here, but the part that is not explained will go into the residual.

So, we characterize or sort of represent the variance of G S_i as sigma squared you know t which tells me the trend component plus sigma squared u right? So, you know the trends are explaining some variation in the data and the rest is remaining unobserved ok. So, this is just to sort of get an understanding of the regression model.

So, now say we estimate beta 0 and beta 1 from the given data, data set using what is called the Ordinary Least Squares method ok, that is the OLS. Now, if you have not heard of OLS, do not worry about it. We are going to move to spatial regression in the next lecture and we are going to introduce what OLS is and what a regression is. And, then we will adapt it to the space, dimension spatial dimension and then move from there to the causal inference idea ok.

So, say we estimate beta 0 and beta 1 from the given data set using the ordinary squares. The idea is that I have the true parameter beta 0 and beta 1 which were a representation of what I should be expecting. Now, I have estimated them, I call them beta 0 hat OLS and beta 1 hat OLS. This beta 0 hat and beta 1 hat OLS are nothing, but values that I am able to reduce from the given data set right?

Beta 0 hat will be the intercept, where it touches the line right; touches the Y axis, the regression line, and the slope at which it is rising, that slope is beta 1 hat right? So, it will be an actual numerical value right? So, once we get our beta 0 hat and beta 1 hat, we are able to construct a mean and trend filter for our groundwater data. What does that mean? Well, once I have my beta 0 hat and beta 1 hat, I can put that back into the regression model that is to say G of S_i minus beta 0 hat minus beta 1 hat i.

Remember, this value is nothing but the residual that is remaining when we remove both the beta 0 and beta 1 i effect. So, we have the u_i left, because we have now included actual numbers as beta 0 hat and beta 1 hat and they have come from the data. Well, this value is nothing, but u hat i, and when I say u hat i it means residual at location i.

So, I could have actually written this as u hat Si or I could then also write this as G star Si, where G star Si is the demean and detrended groundwater level variable. What does it mean when I say that I am removing the mean and I am removing the trend? As I said beta 0 is a representation of the mean value of groundwater level, which is an expectation of groundwater level when i is 0 and beta 1 is the trend effect, right?

Now, the point is that when I estimate these things from the data and deduct these from the data, what I am doing is that I am removing the mean, removing the mean groundwater level. And, here I am removing the trend in groundwater levels ok. What does that? Let us figure it out by going back to our graphical representation. So, we will do it on the next slide.

So, let's say I have my data again, you know let us just remind ourselves quickly that I was working with data at locations S_1, S_2, S_3, S_4, S_5, and S_6. And, the groundwater levels at these locations were 2, 2 meters, 5 meters, 12 meters, 8 meters, 15 meters, and 20 meters. So, my original plot looked something like as under. Let us plot this. I have my locations S_1, S_2, S_3, S_4, S_5, and S_6, and on the X axis, I have 5, 10, 15, 20, and 25 ok.

So, I am going to try and do as good a job as possible to plot this to construct this plot to scale ok alright. We have S_5 is where we hit the groundwater at 15 meters and S_6 where the groundwater is at 20 meters. So, this is the true data set and we had drawn the trend effect here earlier, again trying to sort of keep the representation about the same, not exactly the same ok.

So, this is my trend effect. This is my original plot. I am just going to call this my original plot ok. Now, let us sort of plot, draw the detrended and demeaned plot right? So, now, I am going to draw what I am going to call filtered. So, you know filtered groundwater plot; that means, when I say filtered, I mean demean and detrended groundwater levels.

So, let me draw that, try and be as accurate as possible; so, that you can compare these two plots. S_1, S_2, S_3, S_4, S_5, and S_6, on the X axis I have my 5, 10, 15, 20, and 25. Now, after I demean my data, where is my data going to be you know mean that right? So, first of all, I have detrended the data. So, I have my trend like this, when I detrended it becomes flat.

And, when it becomes flat at around the value of 9, that is the G bar, that is the average groundwater depth, that is G bar equals 9; I am going to have my data you know centered at this mean. But, what I have also done is that I have demeaned my data, that is to say, my mean value will now slide to 0 right; because I have taken out the mean, I have taken out the trend.

So, first I have taken the trend, I have flattened it right and then I have taken it from 9 to 0. So, I am going to try and draw my resulting data set like that. So, to do that, I am going to remove, I am going to just shift the axis a little bit on the Y axis such that it is convenient for me. What I am going to do is, I am going to draw 0 slightly upward; just for my convenience. This is just a convenient plot.

So, I have my mean that has now moved to the point which is 0 on the Y axis. And, if you look at the values that they will be after demeaning and detrending will be something as follows. So, it will be somewhere here slightly above that at S_2 at S_3 it is a little bit more, and S_4 drops down, right? This is S_4, S_5 again it is slightly below the and then S_6 is slightly above ok.

So, now I have my values centered around the mean. So, the mean is what is giving me your u_i hat. So, instead of you know G S_i which is on the left-hand side plot, I am plotting G star S_i which is the demeaned and detrended value ok. G S_i minus beta 0 hat minus beta 1 hat i.

Now, when I plot a variogram, when I estimate a variogram, variogram estimation what will it do? It will provide me with a spatial dependence structure for this entire data set and provides a spatial dependence structure or strength at each observed value. What that means,

is that when I observe you know at S_3, when I observe this blue dot what is going to happen around S_3 is that the values will look similar.

But, as I move away from S_3, those values will start to look less and less similar to S_3 and more similar to the mean, the global mean right? Because, if I get too far away beyond the range, then I have no predictive value coming from the observed value at S_3. So, what happens is that when I move to the right of S_3, I have values that are similar to S_3.

But, as I move forward this spatial dependence goes away and it starts to converge with the point which is the no correlation point, no spatial correlation; that is beyond the range, that is estimated from the variogram, right? So, after the range, these values are going to be equal to the mean, because there is nothing to learn from the neighborhood. As I move closer to S_4, well there will be again something to learn and these values will start to look similar to S_4.

Similarly, as I keep moving forward, the data that I have will start to look as shown on your screen right? This is how the spatially dependent data are going to sort of you know, we are going to be able to predict a value S_0 away from the sample location in our data ok. So, if S_0 was somewhere here, let's say it was closer to S_3 and farther away from S_4, then the S_0 at S_0 location the cross will give me the demean and detrended value.

Now, to retrieve the actual value, I will have to add the trend back to this data right? So, I need actual G Si, right now I have G star S0 right? So, I have G star S0, I want G of S0. So, I am going to go and back and add this component which I had removed from the data ok. So, the as a next step, I am going to say add trend. So, we had detrended then because detrended data is stationary, we estimate our variogram, and we get our G star S_i.

And, then we add the trend back to the data. So, if I do that, I am going to use a different color. Here, I have again S_1, S_2, S_3, S_4, S_5, and S_6. So, what I am doing now is I am adding the trend back to the data. So, I have my trend line, here like this I am going to have to just add the trend as well as the mean to the data. So, the data will look like the following.

So, I have my trend added back to the data, let's say this is the trend added back to the data. Then, I have my representation will sort of change to the following. So, I am just drawing these things to you know not to scale really, but you know just to give you an idea of how these things would look like if you were to work with a real-world problem like this ok. So, if

you were to conduct this prediction, then you can go between S_0 and S_3 and S_4, let's say to S_0. And, now you know you will have your unknown G S_0 estimated from these data.

Now, let's say we did not conduct this detrending effect, let's say we did not do that right? What will happen then you know, is we will work with a demean data set, of course, you know the data are going to be centered around this G_0. When they are predicted, what is going to happen is that the variogram representation will look like the following. So, I am going to use green ink for that. Now, this green ink is representing the scenario when I did not demean and detrend my data ok.

So, I hope this lecture was you know fun for you and I look forward to having you in the next module which is called spatial econometrics.

Thank you.