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

Lecture - 20A
LISA statistics

Hi everyone. Welcome back to the 20th lecture on Spatial Statistics and Spatial Econometrics. This lecture will be the last in the series while we transition to tutorials on Arc GIS and R. And today's lecture, we will talk about two broad concepts, one is the LISA principle and spatial dependence or test statistics called the Moran's I statistic, and the Geary C statistic. And then finally, we will introduce a little bit of hypothesis testing in the spatial regression models, right?

So, to do a bit of recap, we started this course by looking at a lot of spatial data, and sources, including the popular ones that are being applied for various contexts. And then we spent some time developing the idea of spatial dependence formally using statistics, like the variogram, and the covariogram, and then we came on to this understanding of spatial interpolation, right?

In this journey, we have also figured out the importance of spatial stationarity without which none of the statistics that we have seen till now, including the regression modeling, make sense if our data are spatially non-stationary, right? So, I really encourage you that going forward as you apply these tools in your respective research problems or any type of setting you to apply them, you should give a very hard thought, look at your data, and your context, and worry about stationarity, right?

If you think that you cannot perfectly argue out stationarity, but still you have some ideas of where, and in what conditions the data might be stationary or might not be stationary, those need to be seriously documented whenever you produce such results, right?

In this journey, then we have looked at this idea of spatial regressions, and we made this transition from regressions as a statistical tool to a causal inference tool, right? Therein, we mobilize this idea of reflection problem due to a seminal paper by Charles Manski, right?

And from there we have then you know seen Luc Anselin's sort of a school of analysis using spatial weights and spatial weights matrices to conduct spatial regression analysis to model

spatial dependence through either regression error terms, or as a mean outcome term in the regression model, right? Today, we are going to continue that line of thought and we are going to look at these univariate statistics that you can look at.

You can see them as a counterpart to the variogram model or the covariogram model, and think of them as an alternative to detecting spatial order correlation in data, right? The fundamental difference between a variogram and a statistic, for example, on your screen, you are looking at this name called the Moran's I statistic which is a popular statistic to detect spatial autocorrelation is the dependence on your weights matrix, right?

The dependence on a row standardized weights matrix is what differentiates the Moran's I statistic or the Geary C statistic, from what we looked at the variogram statistic which is a fundamental tool to document, detect, and measure spatial dependence in data, right?

So, let us begin. So, we are going to talk about detecting local spatial autocorrelation based on neighborhood weights in the first part of this lecture. And here, this lecture can be seen as divided into 3 broad parts. The first is called the LISA principle. LISA is an abbreviation for Local Indicators of Spatial Association, and we will go over this in detail in a minute.

And then we will look at two popular statistics called the Moran's I statistic and the Geary C statistic, right? I will just write out the overall aim of this part of lecture 20.

So, here the first aim is to identify the location of spatial clusters. And the second is to assess the extent or degree of spatial order correlation in data and comment upon the statistical significance properties, right? I mean statistical significance of what? Of the measure of a degree of spatial order correlation, that is going to be either the Moran's I statistic or the Geary C statistic.

So, first, we are going to talk about the LISA principle, right? So, the LISA principle, as remarked earlier, a LISA is a short form for Local Indicators of Spatial Association. Now, they deliver a local spatial statistic for each location. So, we can say location-based spatial statistics. So, I am going to say that, I am going to modify my last sentence and make it a little bit more concise. I am going to say they provide us with location-specific spatial statistics, alright.

And the sum of these location-based statistics, I am going to say LISA statistics provide a measure of global spatial dependence or spatial autocorrelation in data.

So, we are trying to make a comment upon globals and provide a global statistic. But when we do that we are going to somehow through this LISA principle, we are going to have these location-based spatial dependence statistics that we will then aggregate over space. And this link between local and global statistical understanding of spatial autocorrelation is where this LISA principle is situated.

So, by LISA principle we will be able to say that the global statistic, when I say global statistic I mean global statistic for spatial order correlation is going to be equal to K times, K is a constant, summation of LISA i. So, a statistic that we evaluate, you can say a component, a local statistic at location i, right? So, local meanings location i based, local statistic for spatial order correlation. So, we will be working with these location-based statistics.

We are going to fundamentally define location-based statistics. And then, we are going to define them in a way, so that we can then sort of aggregate them with a constant multiplier and formalize a global statistic. So, that is about the LISA principle.

And now two popular statistics based on this principle are the Moran's I statistic and the Geary C statistic. Now, Moran's I is a more recent, statistic more popular statistic. I am going to spend a little bit more time on that and give you just a gist of the Geary C statistic which is based on the same principle.

So, what we discussed for Moran's I, quite a bit of it will apply to Geary C, but we will also see how they are different, and some broad introductory details. So, Moran's I statistic, if you want to read it read more about it, as I said it is based on spatial weights. So, it comes from the Anselin, Luc Anselin school of thought, right? So, Luc Anselin was you know has pioneered.

This idea of using spatial weights for or weights matrix-based measure or measurement of spatial order correlation, be it for finding a global statistic that we will see now or for modeling through spatial regressions that we have seen in the past couple of lectures.

So, Moran's I statistic is based on two components. So, two fundamental components of the Moran's I statistic are, first a row standardized weights, that is w_{ij}, something that we have seen now, and by consequence and a weights matrix.

A weights matrix, I am going to say w tilde because previously we saw w is a weights matrix which has a binary 0, 1 characterization between whether you know the row representation i are neighbors with column representations j. And the row standardization basically normalized for the degree of interconnectedness for these spatial units, right? So, we are talking about row standardized weights and a weights matrix entering Moran's I statistic.

Now, I am using, when I am introducing, I am saying w tilde. But you know I, as I go forward just for clarity if I use w, I am basically talking about row standardized weights. I am never going to talk about non-row standardized weights. And why is that problematic? For that, we have to go back to previous lectures and then figure that out.

So, the second component is the deviation from the mean. That is simple, right? So, I am going to say this is Z_i. So, deviation from the mean and a spatial weight are two components that comprise a Moran's I statistic.

Then, Moran's I at location small i, is denoted as under. So, I have I at location i. So, capital I represent Moran's I, then at location i is equal to summation j w_{ij}, I say tilde, so weight row standardized weight Z_j Z_i divided by i Z_i squared.

And of course, this location i represent, let us say if I had a domain of interest D, right if I have a domain of interest D and every location, if it is a geostatistical domain every location here is a representation of i. So, if I have a data set that goes from i to 1 to n, then I_i is defined for all those locations, right?

Now, given that Z_i is or represents, let us say represents the deviation from the mean. We have something like an underlying process x, such that Z_i equals x_i minus x bar. So, that is what you know deviation from mean really means. And so, I can rewrite I_i as summation j w_{ij} x_i minus x bar x_j minus x bar divided by summation i x_i minus x bar the whole squared. I am again going to say w_{ij} is equal is the tilde. So, you know we have this entity, ok.

Now, if you just stop for a minute and draw your attention to this last expression for I_i; if we did not have the weights w_{ij}, right? The numerator is a representation for covariance between

you know x at location i and location j and in a denominator, I am basically normalizing it with the variance of $x_{i's}$, right?

So, I am in a way writing a correlation metric. So, a correlation would have been rho summation x minus. So, if you look at, if you think about the traditional covariance correlation metric, then you have rho x y will be summation x minus x bar y minus y bar divided by summation x minus x bar squared, a y minus y bar squared and the whole thing square root.

When we brought this to the spatial setting and we defined what was called the correlogram. In that scenario you know, so if I say rho spatial you can go back and look at the correlogram there it was C x y over C 0, right? Now, if I remove the weights all I am looking at is a correlation metric, right? A traditional correlation metric enhanced by spatial weights is Moran's I statistic. It is as starkly similar as that.

So, let us, with that thought develop this concept further. So again notes, we have said this now x_i, I am going to rewrite the Moran's I statistic for clarity. Let us rewrite it on this page and then move forward. So, I have I_i. I am just writing this out from the previous slide summation j w_{ij} tilde x_i minus x bar x_j minus x bar divided by summation i x_i minus x bar the whole squared.

Now, notes x_i minus x bar in the numerator. So, I am talking about this term here, right, xi minus x bar in the numerator can be taken out of the summation operator, right? Why is that? Because the summation is over j. So, with respect to j, x_i is a constant. So, I do not really sum x_i once I fix a location i, x_i minus x bar is a constant, right? So, that means, this term can come, right out of the summation operator, right?

And if you look at the numerator, it is itself a constant. It is just a variation, sample variance times N minus 1, right? So, the denominator is a constant because it does not depend on i, you are integrating i out of the denominator, right? So, I am going to say that.

So, the first note was that. The second note is the denominator is a constant, right? That is what I am trying to say that it does not depend on i, one very hard to see, right? So, now, we will rewrite our Moran's I as x_i minus x bar divided by summation over i x_i minus x bar the whole squared times summation j w_ij tilde x_j minus x bar. This can be then further written as

x_i minus x bar is Z_i that is the deviation from the mean, summation over i Z_i squared summation j w_{ij} tilde Z_j.

Now, clearly, this term sitting outside the summation j term is a constant, we can say a constant A. So, then I can write this term as A times summation j w_{ij} tilde Z_j. And as I have mentioned earlier, the LISA principle; I have applied the LISA principle, I will now link the local statistic with a global statistic. So, let me do that.

So, I am going to say summation i, I_i will be equal to N times I, which can be seen as a N I. So, as if I am summing N times I's at each location. So, if the local correlation was a constant at each location, I can think about the global average being capital I and N times that global average is going to give me the sum of each individual unit providing me a spatial correlation matrix that is the Moran's I.

So, this would mean that I can be written as summation i I_i over N. And this I is termed as the global Moran's I statistic. And clearly, the global Moran's I statistic is an average of the local Moran's I statistics evaluated at locations i in our sample data. So, now we have sort of developed this global idea of a global spatial autocorrelation metric that is the global Moran's I.

And of course, once I have a statistic, the next step for me is, what about statistical inference, what about the confidence intervals on this statistic. So, that is the next natural step.

The next natural step is inference on or for Moran's I statistic. So, I have I_i which is equal to xi minus x bar over summation i xi minus x bar the whole squared summation j w_{ij} x_j minus x bar. Now, the next step is to define a Z score, I am going to say Z_i score, right?

So, the Z score is for constructing the 95 percent confidence intervals because it's Z for every i, because I am looking at Moran's I at location i. I am calling it Z_i score for the local Moran's I statistic I_i. Z_i is going to be equal to Ii minus the expectation of I_i divided by the standard deviation of I_i which is the variance of I_i square root.

Now, this is a standard definition for any statistic for which we want to you know conduct statistical inference. The question is, how do we construct these entities expectation I i? Where do these come from? Right. So, to construct them we have a computational module,

right. So, I am going to say expectation of I i and the variance of I i are computationally evaluated, ok. So, there are several steps. So, we I am going to go over these steps.

The first step is to hold value or location fixed at location i. So, I have a data set, ok, I have a data set, here is my data set, ok. There are the locations i, j, k, l. So, these are the locations for which I have data observed over an irregular lattice, ok.

So, the data are irregularly spaced and all that. And each location that I observe the data can be thought of as location i. So, I am going to go at. So, for example, if these were groundwater values, then every i, I am going to observe groundwater values.

So, what I am going to do is, I am going to go to location i, and I am going to fix the value of the groundwater level at that location. That is step 1. What is step 2? Step 2 is randomly permutated random permutate or permutes remaining N minus 1 values, right? What does it mean random permute N minus 1 values? So, I have fixed the value at location i, right?

For all other locations, I am going to now randomly assign the values for the remaining N minus 2 values. So, I have my bag of N values at different locations. I am going to pick one and hold it constant. The remaining N minus one in my bag, I am going to randomly assign them to all the other locations. So, it is kind of a Monte Carlo. It is a simulation that is what I am trying to do, right?

So, say I have a sequence of values you know 2, 4, 6, 8, and 10, let us say I have 5 locations, data for 5 locations and this is the bag of values I have. Let us say I go to location 3 and I fix the value at 6. So, I go to location 3, and I fixed my value at 6.

Now, for the remaining, I have values 2, 4, 8, and 10 sitting in a separate bag, these are the ones that I have not fixed. I have not held them fixed. So, what I am going to do is now with the replacement I am going to randomly just assign these values at different locations. I am somehow simulating a process of generating the data. Let us say this is one such construction. With this new data set, now with the new sort of spatial setting what I am going to do is I am going to compute I_i. So, I am going to compute I_i.

And now this is the activity I will do for every location i. For every location i, I go in, I hold the value fixed and then I randomly assign the remaining values to all other locations and then I compute my I_i. For each permutation, each permutation means location because the

permutation is tied to location, right? So, when every time I permute, I am basically at a location i, i plus 1, i minus 1, i plus 2, and so on.

Now, the third step is to repeat steps one and two. So, I am going to say 1, 2, 3 m many times, I am going to say m equals 500 times. Let us say I repeat these steps 500 times. I have really simulated the process of spatial construction in the data in every possible way, right?

And step 4, I will get the mean and standard deviation of $I_{i's}$ drawn from 500 iterations. The set of $I_{i's}$ now I have for each of these iterations, I am going to collect them, get the mean value, and get the standard deviation, right? And also, I can draw confidence intervals in terms of the 5^{th} and the 95^{th} percentile values.

So, basically what I have done is that I have gotten this mean, this mean relates to this expectation I_i, this standard deviation here relates to this variance of I_i square root, right? But I have not gotten to them analytically, rather I have gotten them computationally or by using simulations, right?

Now, all of this is done on the software. We do not really do it. And in the next module, we are going to transition and start doing these things, right? What is important is to know what the software is doing behind the scenes. So, if we see some results which we feel are something that we did not expect, then we can at least try and address those things.

So, the output on Arc MAP or Arc GIS, something that we are going to start as the next module after this lecture, is going to sort of provide me with what is called a cluster and outlier analysis of Moran's I based on a 5 color scheme.

It is going to give me, once I calculate Moran's I the output will have 5, a 5 color scheme, right? And that scheme will have 5 categories basically in different colors, it will say not significant, high, high low, low high, and the next you can guess is a low category, right?

And it will say when it says high and low, it is going to call it a cluster, a cluster that agrees with the positive spatial correlation of nearby values, right? And high low and low high are going to be identified as outliers, and some of those are going to be not statistically significant.

So, the Moran's I, globally is not going to provide me, you know I have a global statistic. But locally you may not have spatial correlation or spatial dependence in data in some regions

and it may have it for other regions, right? So, this kind of distinction is directly feeding, I will say feeds directly into the tutorials on Arc GIS.

And the last thing, I want to talk about in this part of the lecture is the Geary C statistic. Now, the Geary C statistic looks pretty similar by the way to Moran's I statistic. It is given by N minus 1, capital N is the size of the data. Summation i summation j. So, instead of just summation j, we have summation i and j. Row standardized weights x_i minus x bar, the whole squared divided by 2 summation j w_{ij} summation i x_i minus x bar the whole squared.

Now, we know that w I_i is equal to 0 for every i, right? And the interpretation, I am going to provide the interpretation here is that Geary C is inversely proportional to Moran's I. So, that means, if Moran's I provides a signal of a positive correlation, then Geary C is going to provide a lower value. So, if Moran's I is higher, Geary C is going to be lower, right? C_i can is always greater than 0, but it is unbounded on the higher side, right? So, it can technically go to infinity.

When C_i is less than 1, that is to say, that C_i is between 0 and 1, then it represents increasing positive spatial autocorrelation, positive increasing spatial order correlation as C_i is decreasing. And whenever C_i is greater than 1, that means, C_i is more than 1 and then it will represent the situation of increasing negative spatial order correlation as C_i is increasing.

So, as C_i increases overall you can think about it as a measure of decreasing positive spatial order correlation. And, when C_i is greater than 1, you can think of that situation as one of negative spatial order correlation.

Some notes of caution. First, both Moran's I and Geary C are exploratory statistics. We can at best talk about correlations. We cannot talk about causation. We cannot say that an entity in the neighborhood is causing behavior at the location of interest. No, they are just associates. So, they are merely explorations. And both these statistics assume that the underlying processes are the same. So, that is they assume stationarity.

And the way sort of this has been interpreted is that they assume that a given data set, given data set is in a spatial equilibrium. So, in a way that there are no transitory processes in the underlying data that we are looking at. So, if we are looking at land use transitions then this statistic can be problematic. And finally, something that I have already said is that these statistics are strictly univariate. So, they are univariate only.

They have no utility in multivariate analysis. That you have already seen with spatial regressions.

So, with that, this is an introduction to the Moran's I mainly. I have given you some notes of Geary C. Moran's I, is sufficient for our purposes in this course and for most applications, and it is a far more popular statistic today. It is based on the spatial weights. So, if you do not have a definition of spatial weights, you cannot really get to a Moran's I.

And I am ending this lecture at this part of the lecture with some cautionary notes. And I want to sort of drive a point home that we have more sophisticated tools at our disposal in terms of the variogram models to infer upon spatial order correlation.

So, perhaps even though this tool is available, we have a more advanced or fundamental understanding of how to measure a spatial order correlation for very sophisticated surfaces.

With that, I am going to end this first part of lecture 20. We will have a very short, second part where we will introduce hypothesis testing for spatial regression models. And then we will move right on to the tutorial.

So, thank you very much for your attention.