Spatial Statistics and Spatial Econometrics

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

Lecture - 9B
Spatial Autocorrelation Implications for Inference - 2-D example

Alright, So, we will now move on to a second practice problem or a second-class exercise.

So, here we are going to sort of transition from a one-dimensional real line you know the representation of spatial locations to a two-dimensional you know representation of a circle. So, I am going to read down and read out the problem and then we will sort of step-by-step start to resolve the issues that arise with spatial dependence structure in a two-dimensional space.

So, consider a real estate market for a circular city with N equally spaced homes that are located along the city's circumference of K units ok. So, what we have that we are talking about is a circular city; so, I am going to draw a circle alright. And now, we have N equally spaced homes; so, the circumference of this circle is K; so, you know and there are N homes which are equidistant from each other.

So; that means, that there are going to be homes or houses placed on this circle at a distance of K by N along the circumference you know as we go. So, let us call this location 1; so, you know that the first starter I have drawn is location 1. And I am going to go in the clockwise direction and keep sort of adding a dot or a location for a home every K by N units apart ok.

Of course, I am going to do it with approximation, because I do not know N, I do not know what is K by N it is just a schematic not to scale representation on your screens. So, say this is home 1, home 2, equidistant to 2 is home 3, then we have 4, we have 5 and 6, and so on and so forth till we get to our point where we have a home here called nth home, the home price at each location i is denoted as P_i.

So, the price of a house is the market that is kind of observed in the market in real estate market of this circular city is you know indexed by its location. So, at location 1 the price is P_1, at location 2 its P_2, at location 3 its P_3, P_4, P_5, P_6, and all the way to P_n. Now, as these prices are assumed to be correlated with prices of homes that are i's first order neighbor's ok.

So, at any location i, i look at the location on the right to it alright ok; so, one step in the clockwise direction one step in the anti-clockwise direction.

So, for every i, i will look at location i minus 1 and location i plus 1. So, I am going to say for location 2, I am going to say that the price P_2 is correlated with the price at location 2 minus 1 which is P_1, and location 2 plus 1 is a P_3 ok. And, for 3 of course, you know it's going to be 3 minus 1 which is P_2 and P_3 is going to be correlated with 3 plus 1 at location 3 plus 1 which is P_4 right; so, this is how the correlation structure is going on in space.

Now, although this is all quiet you know this is a model and of course, you know it is a representation of what happens in the real world. All it is trying to say is that the real estate prices in the real world can be represented by a situation where you know they are highly correlated with immediate neighbors and not so correlated with those who are farther apart ok.

So, maybe you know you can say that homes in a particular locality can be considered neighbors, and homes in different localities or between localities are non-neighbors right? So, that kind of a situation is what this particular stylized you know setting is providing us ok. We define this set N_i which contains you know the neighbors of i that is N_i is a set that contains i minus 1 and i plus 1.

So, neighbors of 2, neighbors of you know the home or house at location 2 are the house is at location 1 and 3, neighbors of you know the house at location 3 is the house is at location 2 and 4. What about you know the neighbors of you know of a house at location 1 what about location 1, well here the neighbors are n and 2 ok. So, because it's a circular city instead of a real number line where n only had one neighbor; now, it also sort of has a second neighbor which is house 1.

Similarly, the set for you know of neighbors at location n will be location n minus 1 and location 1 ok and so on and so forth. So, we can define this set for to that contains neighborhood houses as N_i, it's a very efficient sort of an efficient mathematical notation I would say ok. So, needless to say for the given case study the N_i number of elements in each set N_i is 2 which is correct.

So, you know we have i minus 1 and i plus 1; so, we have two elements in N_i which is fine. And if j is in N_i, then i must be in N_j that is if j is in the neighborhood's set of I, then i must be in the neighborhood set of j which is also correct if we look at the examples that we have written out here. Let the home prices P_i be drawn from a normally distributed random process such that expectation of P, the expectation of P is the population you know mean mu P.

And the covariance structure is a spatial dependence structure covariance P_i P_j where i and j are locations i and j. So, we have a spatial dependence structure that exhibits correlated prices with first-order neighbors something that has been written down in English right. So, whatever is written in English simply is translated with mathematical you know notation. So, covariance P_i P_j if is sigma square if i minus j equal 0 that is variance at a location i that is the variance of Pi is simply sigma square for all i ok.

All i in 1 to n ok, the covariance of you know of prices between locations i and j; if they are an immediate neighborhood is given by sigma square lambda, and lambda can be between minus 1 and plus 1. So, now, we are allowing for negative spatial data correlation as well in the two-dimensional space right? So, of course, needless to say, the circle represents a 2D world two-dimensional world.

And for homes that are farther apart there is no correlation right; so, we have discussed this now again this is we discussed this earlier. The first query is to write down the full variance-covariance matrix for the given random process ok. So, I am going to have to write down the variance-covariance matrix for the given random process; so, I am going to do that. So, I have been given this vector of price is P which is nothing but P_1, P_2 all the way till P_n.

All of these are distributed jointly right, it's a random function we have working with spatial data something that we have covered in one of the earlier lectures. So, we have to work with a jointly distributed normal distribution right? So, they are jointly distributed normal random variables with mean mu P something that is provided as information, and the variance-covariance given as some sigma P ok.

So, I know that my P is a n by 1; so, n by 1 vector; so, there are n rows and 1 column right; so, P_1, P_2, P_3, P_4, P_5, till P_n right. And so, mu P which is the expectation of each P is nothing but n by 1; so, this itself is a vector and the variance-covariance matrix is going to be a n by n matrix because it should contain all the variance values, and the covariances right.

So, this is something we have seen earlier for a vector that is n by 1 size the variance-covariance matrix is n by n size right? So, the variance of P where P is a n by 1 is given as sigma P. So, I am going to draw a n by n matrix; so, I am going to have n columns and n rows. So, let us say each row represents i; so, i equals 1, 2, 3 all the way till n, and columns represent locations j so far as the covariance structure goes ok.

So, if for the first cell of this variance-covariance matrix when i equals j that is i minus j absolute value is 0, then the variance-covariance, covariance structure will tell me that the diagonal element will be sigma squared. Similarly, for all diagonal elements when i and j are equal exactly equal, what I am going to find is a sigma squared value at all diagonal elements. For the first off-diagonal element where i equals 1 and j equals 2, we can go back to our covariance structure and write this value to be sigma squared lambda.

For the third position; however, 1 and 3 are two positions apart with no correlation we will have a 0. So, we will have a 0 at 4 at 5 and keep going at location n we have we know that in the neighborhood set of locations, 1 n is one's neighbor and is in fact, the first-order neighbor of 1 right? So, 1 will have spillover from and to location n and location 2 ok; so, the value of this off-diagonal element in the first row and nth column of this variance-covariance matrix is going to be sigma square lambda ok.

Similarly, when we move to i equals 2 and j equals 3, I am going to have sigma squared lambda at 3 and 4 sigma square lambda. So, you can see that the elements that are right you know besides this element these diagonal elements, you know sigma squares are going to be sigma squared lambda. Now, the variance-covariance matrix is symmetric; so, you are going to have sigma squared lambda, sigma squared lambda 0 and so on and so forth right?

So, you are going to have sigma square lambda sigma squared lambda ok. Now, just like you have you know n is in the neighborhood set of 1, and 1 is in the neighborhood set of n. So, we will have you know in the spirit of symmetry of the variance-covariance matrix the element in row n and column 1 of sigma of variance-covariance matrix sigma is going to be sigma square lambda.

Everything else is going to be 0s I am going to have all these things are going to be 0s ok. So, this is the first query that you know this matrix that I have written down here is the variance-covariance matrix of the given random process, I am going to write it again just; so, it is absolutely clear. So, it's n by n all the diagonal elements are sigma squares ok, the off-diagonal elements the first order off-diagonal elements are sigma squared lambda ok.

And then the second-order diagonal elements are 0s except for this aberration between one and n which happen to be neighbors. So, I am just going to fill them in ok, I have 0 0 and then sigma squared lambda ok I hope this is clear. So, you should write this out by yourself; so, that it becomes very very clear what is going on here ok. So, now, we have this spatial dependence structure condensed in this n by n matrix sigma; so, let us take the next step alright.

Next, it says say you have a sample meaning realization at each value on that circle where P_1 is 1, P_2 is 2, and so on. So, the value of P_i is simply i which is the value of the home at location i is just i. So, somehow the location itself is an indicator of value right; so, it is a specific again a very specific example. Then deduce the consistent estimator of rho of omega P sorry mu P which is the population mean.

And those for the 95 percent confidence bands of mu P, how does your estimator depend on spatial heterogeneity if any in the given data, very carefully we are talking about spatial heterogeneity something that you have learned heard of and learned earlier ok. So, again I am going to quickly write down you know just visualize my data because this will help me you know move forward P_1 to P_n it turns out that P_1 is 1, P_2 is 2, P_3 is 3, P_4 is 4, P_5 is 5, P_6 is 6 and so on and so forth and P_n is n ok.

They are asking me what is the estimator of mu P, well we have learned earlier the best guess of mu P from a given sample is just P bar which is summation i equals 1 to n P_i by n and this is going to be nothing but summation i equals 1 to n i over n ok. Interestingly in the numerator, I have the sum of you know n natural numbers going from 1, 2, 3, 4 till n right. And you know you must be aware from high school that this value is n n plus 1 over 2.

And there is an n sitting in the denominator; so, I have n plus 1 over 2 this is n plus 1 over 2 is the sample mean, and my best guess for what you know the population mean would be ok. Now of course, the P bar itself is comprising P_i which is a random variable; so, the P bar is a random variable. If it is a random variable and I have a point estimate of the P bar in terms of n plus 1 over 2 it comes with an error because random variables can take multiple realizations right.

So, the idea is that if it comes with an error how do I contextualize that error or come up with a 95 percent confidence band to have some level of certainty of where what is the range of values that you know mu P can take as far as it is represented by P bar right. So, to evaluate the confidence band of mu P right we need an estimate of the variance of P bar ok.

And the variance of the P bar is nothing but the variance of summation i equals 1 to n P_i over n. And we saw this in a couple of lectures ago that when you have spatial dependence we can represent this variance as a covariance you know formulation which is nothing but 1 over n square summation i goes from 1 to n summation j goes from 1 to n covariance P_i P_j ok.

And now to be able to sort of sum these covariances for all different you know combinations of i comma j, remember i can take value 1 and j 2 and vice versa i can be 2 and j 1; so, we have to double count these things right. So, what we are doing here when we are evaluating this double summation operator is we are counting all the elements in the variance-covariance matrix.

So, this first step was indeed very helpful, right? So, this will basically what I take doing is I am summing sigma square, sigma square in the diagonal elements n times. For the off-diagonal elements I am summing 2 n times because you know they appear twice on the right of the diagonal and the left of the diagonal ok. So, I am just going to follow that I am going to say 1 over n squared I have n sigma squares and 2 n times sigma square lambda right?

And so, I have sigma squared over n 1 plus 2 lambda ok. Now, if P were spatially independent lambda would be 0 and I would come back with I will get back my sigma squared by n valuation. So, by this spatial dependence structure and the degree of spatial dependence that is exhibited by this parameter lambda, there is a bias that is induced in the variance if we were to ignore the spatial dependence in these data ok.

So, if I have my variance of P bar, I can evaluate my standard deviation of P bar which is sigma over root n times 1 plus 2 lambda the square root ok. And this will imply the 95 percent confidence band for you know mu P can be given as P bar minus 1.96 sigma over root n times 1 plus 2 lambda ok comma P bar plus 1.96 times sigma n times 1 plus 2 lambda.

Of course, 2 lambda has a power of one half which is the square root of this value; so, this is the confidence band 95 percent confidence band of mu P. And you know; obviously, we know that the P bar is nothing but n plus 1 by 2. So, these things are this is something we have evaluated right you know on this page. Now, the next question the last part of this question says how does your estimator depend on spatial heterogeneity?

Now, we had said that spatial heterogeneity is a pattern that represents a long-term trend. For example, when we were looking at you know the real estate values across Delhi, we saw a low to high gradient from you know the west to the eastern part of the city ok, or the national capital territory of Delhi, right? So, what that would translate is that somehow spatial heterogeneity should be a function of i right? it should be a function of location.

And if you look at both the P bar and variance of P bar and by extension standard deviation of the P bar and the confidence bounds none of them depend on i, they are all independent of i right? So, what we see here is that the P bar is independent of i that is it does not depend on the location, and it is not varying by location ok. Higher-priced homes are not clustered at a given sort of location on this circle relative to lower-priced home at a different clustered location right, we do not have a P bar as a function of i.

Furthermore, the variance of the P bar and standard deviation of the P bar is independent of i and by extension, the 95 percent confidence bounds or the confidence interval for mu P is also independent well. It is not mu P it is supposed to be P bar because mu P is a truth it's constant right; so, the confidence interval will be for P bar; so, that is something you can call a typo in the question, but I have corrected it here ok.

So, is also this 95 percent confidence interval independent of i; that means, that you know our estimators do not depend on spatial heterogeneity there is no impact of spatial heterogeneity on our estimators right? So, spatial heterogeneity emerges due to trends that are a function of i; none of our estimators are a function of i, hence there is no impact of spatial heterogeneity in these data alright, so, let us move forward.

The next question is to calculate the amount or the quantum of bias generated from not accounting for spatial dependence in these data right? So, what is the amount or quantum of bias that would be realized if we did not you know account for spatial dependence when the data is indeed exhibited in that structure ok? So, we are given that n is 100 and you know lambda can take various values from negative to positive right, where I started this lecture I said you know lambda can take positive and negative values.

So, of course, we can solve this problem for you know each case, but we can also solve it for a general case. So, first of all, you know the P bar remains n plus 1 over 2 with iid and spatially dependent data ok. That means, no bias irrespective of lambda value lambda could be 0 or it could be a negative or positive value between minus 1 and plus 1 right, it will have no impact on the P bar.

Second bias is the variance of the P bar, well there will be bias and we saw on the previous slide we can go back and check where we said that in case of you know spatially independent data the variance of the P bar would be sigma squared over n. Whereas, here it is sigma square over n plus sigma square over n times 2 lambda right.

So, we have sigma squared over n plus 2 lambda times sigma square over n which is the spatially dependent case minus sigma squared over n which is the spatially independent case, and hence the bias is two lambda sigma squared over n right. And of course, you can now you know put in you know lambda n, and you can you know get these values. Similarly, you know bias in the standard deviation of the P bar which you guys can calculate quite straightforwardly. And then you know finally, the bias in confidence bounds will come from the bias in standard deviation, right?

So, bias in confidence bounds, well more specifically 95 percent confidence bounds will come from will originate from bias in the standard deviation of the P bar. I am going to write this value and I am going to let you know verify it I am going to let you verify it. So, this bias when you eventually verify it will come out to be 2 sigma over root n 1 plus 2 lambda the power half minus 1 whole multiplied by 1.96 ok.

Now, of course, you know we have asked for these values at different levels of lambda. So, you know I can just provide this value for the case for the variance we have lambda which is which has minus 0.7 minus 0.2, 0.4 and 0.9 right. And the bias in variance of P bar will be correspondingly minus 0.014 sigma squared minus 0.004 sigma squared 0.008 sigma squared and 0.018 sigma squared.

So, I am writing corresponding bias values you can go back and check you know at your any your time ok. Similarly, the bias here with lambda will be lambda is all these values minus 0.7 minus 0.2, 0.4, 0.9 right. And bias in standard deviation of the P bar will simply be you know the first one will be a complex number because the square root of a negative entity.

So, minus 0.1 plus 0.06 iota times sigma comma minus 0.02 sigma 0.03 sigma and 0.07 sigma ok, I strongly encourage you to go ahead and verify these solutions. The bias in the 95 percent confidence interval as I have said will simply originate from the standard deviation of the P bar; so, there will be simply these values right; so, you can simply carry them forward to this case ok.

Let us move forward to this circular city example and you know we finally, have this query where we are saying that what if we were to extend this first-order spatial dependence structure to a second-order spatial dependence structure? Not only that where the strength of interdependence among the first-order neighbors will be double the strength of interdependence among the second-order neighbors.

So, now we are sort of relaxing this restriction that spatial spillovers or spatial correlation only occurs with the first-order neighbors, we are saying that well it can also occur with second-order neighbors ok. So, we have a circle again we have our P_1, P_2, P_3, P_4, P_5, and P_6 all of these are equidistant my figure is not to scale. So, you should not sort of you know if some pairs look closer they are not closer you know they are all equidistant ok, this case is slightly more interesting.

It says that you know there will be a spillover on the value of P_2 from P_1 and P_3, but also a spillover from you know P_n and P_4. Although, the spillovers that are represented by these yellow arrows which are from the second-order neighbors are half the strength or half the strength of the first-order neighbors ok. So, let us try and exhibit this problem first of all in the variance-covariance matrix.

So, we have solved this problem for the previous case and we saw that you know the variance-covariance matrix is really good in encompassing all these you know knowledge. So, I am going to write down the variance-covariance matrix for second-order neighbors, the variance-covariance matrix for second-order neighbors ok I mean second-order neighbors and first-order neighbors.

So, we have neighborhood spillovers off the second order right; so, we are not just saying yeah only second order you know neighborhood spillovers that would be quite unrealistic you know setting to work with. So, here rho sigma P is nothing but you have your diagonal elements which are simply sigma squared, sigma squared, sigma squared keep going all the way to sigma squared ok.

Now, you have on the off-diagonal elements you have sigma squared lambda and in the second order of diagonal element you have sigma squared lambda by 2 ok. And if I keep going forward if you will realize that 1 will also have a spillover from location n minus 1. So, you will have sigma squared lambda by 2 and you know sigma squared lambda to top it off all others are 0s ok.

Similarly, I can go ahead and fill up these you know these entities sigma squared lambda, sigma squared lambda, sigma squared lambda, I have sigma squared lambda by 2, sigma squared lambda by 2 right, and sigma squared lambda, and sigma squared lambda by 2 and so on and so forth right. You are going to have for entity 2 you will have sigma squared lambda by 2 sitting here everything else will be 0s right.

So, you sort of get a gist of how these things will you know look like. So, you have sigma squared lambda, sigma squared lambda by 2 this is for the nth entity. You know you have sigma squared lambda, sigma squared lambda by 2 and all other are 0s ok. So, you can sort of; you can sort of you know fill these things in you know, and construct a variance-covariance matrix.

Again, the estimator for mu P is simply the P bar right; so, again the estimator for mu P is simply the P bar which is n plus 1 over 2 right, but the variance of the P bar will be different right? So, you know now that the variance of P bar is 1 over n square summation i summation j covariance P_i P_j. And what we said earlier is it is simply this entity double summation operator is simply summing all the elements in the variance-covariance matrix sigma right?

So, we are simply summing all these elements in the variance-covariance matrix ok. And if you sum them what you are going to have is 1 over n squared sigma squared appears n times, sigma squared lambda appears 2 n times, and sigma squared lambda by 2 also appears 2 n times. So, now, your bias is going to be slightly more than it is going to be 3 lambda.

So, in case when there is no spatial dependence it is sigma squared over n; in case you know in the case of first-order dependence we had 2 lambda for first-order dependence structure right? And 1 plus 3 lambda is for the second-order dependence structure. So, as spatial dependence increases the bias in the variance of the z bar and consequently in the standard deviation of the z bar you know the confidence bounds of the z bar will start to increase which will the bias will enhance.

So, you will be penalized more in terms of your estimators or your you know the precision of your estimators, if you start to ignore these biases or this spatial dependence structure ok. So, this was a very interesting example that sort of transitioned us out of that one-dimensional analysis of spatial dependence structure and it is the implication it is consequences to a two-dimensional structure right?

So, we started with a simplistic understanding we came to a little bit more sophisticated understanding, and going forward we will have an even more sophisticated understanding or you know of these entities. As a next step which we will do as a separate part or portion of this lecture. I am going to talk about Monte Carlo simulations which is a numerical strategy to evaluate the bias in the variance standard deviation and confidence bounds of you know of P bar when you know spatial dependence exists in the real estate market.

And we do not have to rely on these analytical methods which can be mathematically quite complex or analytically quite complex. We can fall back on simulations and we will give you an example of this with the same setting that we are working with right now alright. See you in the next part of this lecture.