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

Lecture - 14A
Spatial Statistics and Spatial Econometrics

Hello everyone. Welcome back to lecture 14 of Spatial Statistics and Spatial Econometrics. In this lecture, we are going to take up from where we left off in the previous lecture which is Variogram Modeling. Up until now, we have defined a theoretical variogram, and we have learned how to develop an experimental variogram which is nothing but applying the idea of a theoretical variogram given sample data. And where we ended our previous lecture was moving from this experimental variogram to a variogram model.

Why do we need a model and what is a model? Well, a model is a generic or a general representation of a real-world phenomenon and it typically relies on certain parameters of interest in the physical environment or the social environment that we are studying. It is only a representation of the real world. Of course, it does not capture all the complexities of the real world, but it will be able to provide us with a generic understanding of spatial dependence in space right.

So, you know we looked at these figures that are in front of your screen, and we were working with Uttar Pradesh data. We were working with you know lag vectors in the east-west direction right? We figured out a way to work or define lag vectors and search for you know spatial pairs even when the data are spread on an irregular lattice right?

And finally, we came up with this variogram cloud which you know provided us with different scatter plot values of z of s minus z of s plus h the whole square, and for each h value that is on the x-axis right, for each h value, we collect all the different points on the variogram cloud and we take a mean right? That mean value is then collected for different h values providing us with an experimental variogram ok.

Now, the point is we want to move from this experimental variogram to a generic understanding of spatial dependence in this region right? What is the first thing that comes to our attention, our attention when we move from this experimental plot to the modeled variogram of you know understanding? Well, we have a variogram measure for specific

values of h which are represented by these red blocks right? So, we know at lag h_1, h_2, h_3, h_4, h_5, and h_6, we know exactly how spatial dependence is exhibited in the sample data set.

But we really do not know what is happening on points between and other than the specific values of h right? But that is sort of limiting right; I mean, if we are to conduct spatial prediction, spatial regression, then we should be able to have a model wherein we provide the value of h and it provides me a measure of or an estimate of the spatial dependence in my sample and spatial domain of interest right.

So, in that spirit, we need a variogram model for all possible h values. So, to be able to exhibit spatial dependence, and provide a specific measure and estimate of spatial dependence, for all h values we need a variogram model. The other alternative is to manually calculate the experimental variogram value for each and every h value in every direction and every distance you know combination that is out there. Now, that is not really possible you know in practical terms, right?

And the other thing that we will see through this lecture is that a generic model will allow us to incorporate additional drivers of spatial continuity, besides just the space location and geographical factors. Sometimes, spatial continuity or spatial contiguity is also driven by anthropogenic factors.

For example, if we think about urban areas and we think about groundwater levels; well, you know because urban areas are typically highly concretized, there is going to be very little natural recharge due to rainfall. Because you know most of the land area is concretized right.

So, if we are studying recharge as a variable of interest over space, then perhaps urban areas will provide a factor that will drive the continuity of low recharge values over the entire urban area. This is relative to other agricultural or non-urban areas, where the recharge value is going to be spatially continuous and typically higher than the recharge values in the urban area. So, anthropogenic impacts are important as well other than just the location indices that we have been thinking about seriously.

An experimental variogram does not provide us an opportunity to characterize such nuanced factors in defining spatial contiguity. It is mostly dependent on the location which is the lag vector h and it is done 1 h at a time right? So, apart from providing us with a generic

understanding of spatial continuity in space, it also allows us to incorporate information that is beside space; but can lead to spatial dependence or spatial contiguity in the data.

In that series of arguments you know a variogram model will also be very useful for spatial interpolation or spatial Kriging. All of these spatial dependence measures that we are working with or you know trying to work out is because you know typically when you sample in space, you know you cannot sample every single location in space.

There are going to be locations, where there are going to be no sample observations, and using what you have observed, you have your job as a statistician is also to predict what might be going on in areas that remain unobserved right? ok.

So, with that let us move forward and study a few variograms you know models. So, first up is the linear variogram model. You have the visual characterization of the linear variogram model. Let us start with the visual characterization.

Well, you have a typical sort of shape, where the model you know variogram values as h increases they go upright. So, on the right-hand side corner, we have this experimental variogram which basically shows that as h increases the variogram values go upright. So, we can see our nugget effect and we can also see that as h goes up you know the variogram values are increasing.

But they are now increasing in a linear fashion that gives us a shape of how the variogram values are progressing as h increases right? This specific format or structure of evolution of spatial dependence as a function of h is where the modeling is coming at ok. And it is coming with another parameter which is the slope of this linear line which exhibits an increase in variogram with 1 unit increase in h right.

This slope and the nugget together provide us a parameter vector theta that is c 0 being my nugget that is very near just when we leave h that is h approaches 0; but it is slightly positive, slightly away from any given location what is the level of spatial dependence right?

That is c 0 and after c 0 b is or b h gives me the rate of decline in spatial dependence in space right? This rate of decline is linear and constant as h increases. So, we do not really get to see this point, where the variogram stops increasing in the linear variogram model. Again, a model is an imperfect understanding of the real world. We know we have seen examples of

how in the real world, you will see the variogram spatial dependence eventually decline and that makes perfect intuitive sense right?

That is values that are farther apart in space are likely to have you know not so strong correlation. So, let us look at the mathematical form of this linear variogram model. So, the linear variogram model is valid in a general d-dimensional real space and you know it is given as gamma h colon semicolon theta equals 0 if h equals 0.

So, this respects the theoretical property of a variogram that gamma 0 is 0 and when h is not equal to 0 that is there is a positive lag between two values, two locations in space; their values will be dependent due to this measure called c 0 plus b l time h right?

This b is the slope at which the variogram, the spatial dependence, and declining or alternatively, this variogram is rising as h increases right? c 0 is the nugget effect, right? c 0 is the nugget effect and you know that when h approaches 0 gamma h theta will be c 0. So, moving away from an experimental variogram or rather a semi-variogram because we are working with gamma h and gamma h theta rather than 2 gamma h and 2 gamma h theta.

So, I am using variogram and semi-variograms you know as interchangeable terms; but I am sure you know that 2 gamma h is called a variogram conventionally, and if I remove this factor of 2 and work with gamma h, I am talking about really the semi-variogram ok. So, here, in this case, you know the experimental variogram gives me values gamma h at each given lag you know in space in my domain of interest, right?

With the model, we have another component of theta. This theta which is the parameter vector is the basic differentiator between the experimental variogram and the modeled variogram. The beauty of theta is that if I figure out these theta values, if I figure out c 0, and b, I can just provide a value of h and it could be any h not just the ones that I have manually calculated my gamma h value for and I can back out my variogram right? So, that is the first strength that comes out right.

So, if I have a parametric definition, if I apply a parametric definition to the variogram itself which sort of in a rather imperfect fashion provides me how gamma h or 2 gamma h will progress as you know as it changes that provides me a modeled variogram or a variogram model ok.

Now, there are several models for a variogram, to sort of you know provide a generic understanding of spatial dependence in space, there are several versions. Linear is one; perhaps the most simplistic a very good point to start teaching these methods. The second on your screen right now, the second kind or second type of variogram model is called the spherical model. The spherical model is valid in one-dimensional, two-dimensional, and three-dimensional you know real spaces ok.

Now, this is not this is to basically say that the linear model was valid in a general d-dimensional real space. Here, if we are using a spherical model, our data better be you know in three dimensions or less right, which is powerful enough right? I mean most of the work that we will do especially for me as a social scientist, as an economist, you know most data sets would form would fall in the category of either in three-dimension and mostly, in two-dimensions ok.

So, what do we have now? Now, we have a situation, where you know we again have our nugget effect right? After we move from a given location that is the origin, you know right after we move a step delta, we get a certain value of gamma h which provides me an idea of the nugget effect.

After we reach the nugget effect and we still keep increasing our h value which results in an increase in gamma h value as well, which again follows really well from what we have learned from the experimental variogram and as well as from what we have learned from the textbook variogram theory right?

But the spherical model can also provide a pretty realistic picture of what we saw with the experimental variograms and the textbook variograms, that is there will come a point after which I am going to have my variogram value stabilized to a given point, right? So, the spherical variogram provides us with a much more realistic picture of what is happening in the real world.

Now, let us come to the mathematical formulation of the spherical variogram and focus on the parameters that allow us to move from the experimental or the calculation of a variogram, manual calculation to a variogram model, which is more generic such that we can simply provide the value of h pass retain the model and will provide me an estimate of what 2 gamma h should look like ok.

Now, again, I have my theoretical property gamma 0 equals 0 satisfied. So, thumbs up. Second, I have c 0, for h between 0 and a value a s it progresses with c 0; c 0 being the nugget effect plus c s times something in a curly bracket minus half something else in another curly bracket ok. It is a spherical variogram. So, you see that h has a cubic form, you know and it is declining at a cubic rate.

So, there is something that is declining if h when h increases at a cubic rate in this spherical variogram model. My theta space is now c 0 c s and a s right? Now, if I pay attention h after h greater than a s is a constant value c 0 plus c s, now this is very close to what we have learned till now. After h crosses a value of a s right, gamma has a fixed value of c 0 plus c s right, c 0 plus c s.

Now, this c 0 plus c s is what we know is our sill. So, what is c s? Well, c s is what we called as the partial sill. What is a s? a s is what we called as the range that distance in space after which any two locations will not exhibit any correlation. So, this sill is the point, is the level of no spatial correlation in data. This spherical variogram is very very close to the textbook variogram understanding that we have right?

And if you want to sort of visualize how the spherical variogram is really working, well it is measuring if you were to visualize two spheres being put together exactly overlapping with each other and as you sort of start to move these spheres out, the level of interest you know as the area that is the intersection of these that is between the intersection of these two spheres, it starts to decline.

And as this starts to decline that gives me a modeled understanding of decline in spatial dependence that is what this function between h, you know h is between 0 and a s is providing mathematically. It is the decline of the area of you know intersection between the two spheres as they are pulled apart from each other. It is a pretty interesting you know physical interpretation of a spherical model that is why it is called a spherical model.

In that spirit, there are a couple more models which are popularly used. The other one, the next one on your screen is called an exponential model right? It looks much like a linear model, only now there instead of a straight line, I have an exponential decline in spatial dependence on data and this rate of decline is where the exponent is going to come in. So, this exponent a e is going to be a parameter of the rate of decline, how quickly will this decline right?

And as h approaches infinity, you will see that gamma h value will now that approach c 0 plus c e which is nothing but the sill that is the point of no correlation. So, the exponential model exhibits a real-world situation, where values are spatially correlated even though they are very far apart from each other right?

So, those if you are working with a physical or social phenomenon that exhibits that value in your spatial domain right? Remember all this analysis is contained in a box that we call the spatial domain. We have used the notation d for this spatial domain, right?

So, far apart enough would basically mean in this circular spatial domain, you know if values are at two different corners of the circle, they are very far apart and if you still believe, if you believe that they will still exhibit spatial dependence; but that spatial dependence will sort of you know decline very quickly as one moves away from any one location, then perhaps the exponential model is the way to go.

Instead, if you believe that there will come a point after which there will be no spatial, you know relation in these data; then perhaps, the spherical model is the way to go. So, these things really matter depending on the kind of real world, you know analysis that you are conducting with the data alright?

Finally, there is another one called the Quadratic model. Again, its parameters will provide us with where the nugget is how far are we going to reach the sill, and what is the rate of reaching there. The only thing is that once you start from any given location, the decline is increasing at an increasing rate. So, there is a convex function right from the origin of this variogram plot.

So, the spatial dependence you know shoots down very quickly after which it sort of stabilizes and then, it decreases at a decreasing rate alright? So, it becomes a concave function. So, if you have such a physical interpretation, then perhaps a quadratic model will be better than the spherical model or the exponential model ok, alright?

So, now, that we have you know appreciated the value of studying or representing spatial dependence through a model, more specifically a variogram model and we have seen some variants of these variogram models, the next step is to actually estimate this variogram model. And when I say that we are wanting, we want to estimate a variogram model, what we want

to really estimate is the parameters through that parameter vector theta that is the essence of the model right?

So, gamma h is the experimental variogram; whereas, gamma h theta is the variogram model that is the modeled variogram right? So, when we say variogram model fitting, we are basically coming to a point, where we want to sort of you know learn the parameter vector theta from the given data set or data sample.

So, we start with the family of linear variograms, right? So, we start again, we started with the example of a linear variogram because it is very easy to understand. It has only two parameters; one is your nugget and the other is your rate of decline in spatial dependence or the rate of increase in the variogram value as h increases right?

So, the value of linear radiogram is given as the set of 2 gamma such that gamma h is equal to c 0 plus b h; where c is greater than or equal to 0 and so is b h right? So, again, this is really a family of the variogram. Why? Because c 0 and b are both variables; so, as I change this pair c 0 b as I insert different values of c 0 and b in different combinations, I will start to get different variogram models right; linear variogram models right? Model fitting then is the concept, where we are seeking an element from the above family that is closest to my given data sample.

So, the linear variogram model is providing me with a family of variogram possibilities. What I want to do is I want to figure out which among this family of you know which member of this family is closest to my data set. That is I want to fix the values of c 0 and b such that my variogram model c 0 plus b h fits really well with the data. Now, the parametric subset of variograms, which will do that is going to be 2 gamma such that 2 gamma h equals 2 gamma h and e and theta; where theta is in the parametric space right?

So, wherever my experimental variogram will match my modeled structure which is here c 0 plus b times h right, twice of this, wherever this will happen I will simply get my variogram; I will simply get my theta value right? Remember we saw that with h, I have a variogram cloud. So, for every given value h, I have different values that are scattered in space. So, h_1, h_2, h_3 right, h_4 right and so on and so forth.

So, I really have a cloud and I have a modeled understanding of what will fit the mean value of this cloud at each value h. So, I am simply going to equate what this you know at every h1,

I have modeled the value of the variogram which at h_1 is going to be equal to c 0 plus b h_1 right?

So, if 2 gamma which is the experimental value, you know matches that are the family of parametric variograms that we will say are the fitted variograms and the value of theta or b and c 0 that if I can back it out from equating these you know at every value of h, those values of b and h b and c 0 are going to provide me fitted values of the variogram model.

Now, we have to find these thetas with some goodness of fit criteria given a sample of spatial data over a stationary domain; very very important, if you were working with a non-stationary domain, well we will not really be able to define a variogram, let alone estimating the parametric vector theta ok. So, with this understanding, we are starting with the general family of variogram models, you know using the parameter vector; then, we are trying to match it with our data and trying to find a close fit.

Now, in doing so, there are several algorithms available. Most of these algorithms, in fact, all that we are going to look at in this lecture or in this course are canned in software. So, when you actually are trying to figure out the theta value that fits or best represents your data set, well you will do it on software. However, in this lecture, you know we will study the process of getting there right?

These algorithms that are in front of your screen are popular algorithms you know in general, I mean they are not restricted to just variogram modeling. Maximum likelihood estimation, least squares estimation, and generalized least square estimation are all very popular algorithms for achieving a good fit for any model and any given data set.

Here, in this case, we have a data set for 2 gamma h right for every value h, we have different values, and we can actually figure out the pairs for each i comma j pair, you know we are going to have i_1 j_1 and so on and so forth and we are going to have these values, the data that we have for 2 gamma h right?

For these values, we will then try and find the theta value which best represents what is happening with the data right? So, we have to be estimated is this theta vector and the algorithm such that you know fitted values that are after we figure out the value of theta are close enough to the true values that are coming from the data set in general.

So, you know in the next part of this lecture, I am going to you know go over these three algorithms in detail and hopefully after that variogram model fitting will become very clear.

Thank you and see you in the next part.