Lecture - 51

Practical
Spatial Statistics and Spatial Econometrics
With R
Prof. Saif Ali
Department of Social Sciences and Humanities
Indraprastha Institute of Information Technology, Delhi

Session - 05
Lecture - 51
Computing experimental variograms in R - II

Hi, welcome to another practical session, where we are learning Spatial Statistics and Spatial Econometrics using R, I am Saif and this is session 5, and today's topic is Computing Experimental Variograms, Part 2. Last time we did Part 1 and I hope that you all watched that because if you did not watch that then maybe today things will not make sense.

I will not do this whole spiel again, if you have been watching the videos, you have probably heard me say it many times. We are focusing in these sessions on gaining the programming skills needed to perform spatial statistics and spatial econometric operations in R.

But programming skills are founded on clear conceptual understanding and programming skills and understanding are involved in a kind of circular relationship, that if you clearly understand the theory, then it should be easy for you to program. And the more you program, the more you work, you try things on your own, the more mistakes you make, and the more number of times you fix them and go back to trying again.

The more you play with the code and think about the results that you are seeing, that should hopefully clarify your conceptual understanding of the subject. This is one of the best ways to learn the subject is by applying it to real-world problems. And I hope that at this stage in the course, some of you are actually starting to notice that if you are not noticing that your understanding and your skill acquisition are feeding off of each other, then maybe you can review or reassess where it is that you need to spend more time.

So, last time we estimated a variogram, we talked about stationarity, we also talked about how to remove trends and if there is a spatial trend in the data what that means, and how to remove it. This time we will go further in estimating variograms but when we do that, what you should be familiar with, which I will not necessarily really explain is the idea of anisotropy. The idea is that spatial variation is not necessarily similar in all directions. It is

not isotropic; things vary spatially in one direction in a different way and in another direction in another way.

And this is something that we will point out in the data, but I am not going to go into it conceptually. So, this is something I am sure you already know, and also the idea of binning that you have a continuous range of values and then you bin them into some discrete intervals, this is something that we will use. So, there is something we should already be familiar with.

Skill-wise, you should have already estimated an omnidirectional variogram. So, an omnidirectional variogram is where direction plays no role, it is the same in all directions. We are saying that the spatial variation in the data is the same in all directions, which is rarely ever true, but that is what we did in the prior video because it's simpler.

And this is something that you should know how to do because we are going to build it on top of this. Of course, you know the meuse data set already and so what will we do this time? We will add the idea of direction or anisotropy to our variogram estimation and we will play with cut-off and width parameters. We will not do this. So, estimating a variogram using a new data set, is something that we will not do. So, that is a scratch that we will do in the future.

So, this is a term that you might come across in if you read material about geo statistics, or if you read papers there is something called the direction of maximum continuity and that is a sort of big term. But what I want you to do is to understand this more and to also appreciate that spatial variation differs by direction, that you know in certain directions things vary a certain way and in other directions maybe not so much.

So, this is the data set that we have been looking at, this is zinc concentrations in parts per million over the river bed of the river meuse. Now, I want you to look at this plot and tell me which is the direction in which the values are most similar or what is called the direction of maximum continuity of values or conversely the direction of least variation, in which the variation is the least.

So, you can look at these sorts of cardinal directions 0, which means North in the gstat package, and 0 means north which is what I have said here, this is from the gstat literature, which you can click on this tutorial here and go to which is on page 13. So, 0 is North, 90 is

East, 45 is North-East, and 135 is South East and of course, 0 is North but also South, I mean it is the North-South direction right, it is this direction.

So, 0 to 135, you kind of cover everything, everything else like 90 is both 90 and 360 etcetera, right? So, all you need is 0 to 135 to cover the whole range. So, let us begin with 0. So, if you move in this direction, you encounter some circles, right? So, they are of a certain size and then smaller in here kind of medium size then large here, and then again large. So, there is not a lot of continuity, I mean circles seem to vary in size in this direction.

Similarly, with 90 again like larger circles here, and when I say 90, I mean if you draw parallel lines like 90 direction over the whole domain, not just this line. So, if you just move in this direction from left to right, like large circles and then followed by smaller ones. So, even in the 90-degree direction, there seems to be some variation.

Well, how about 45? So, if you move in this direction. So, if you start here and let us see we start moving this way. So, we kind of encounter circles all of which are kind of large, and then if we start here and move this way, then all of the circles in this line are kind of similar in size.

So, it seems that in the 45-degree direction, there is the most amount of continuity that is a vague sort of term, but visually at least there are the most continuous values encountered in this direction. And because spatial trends are not good and we like to have small variations in values when we are estimating variograms, often variograms are estimated along the direction of maximum continuity.

And there is a large discussion around this, this is not again, there is no hard and fast rule for why you have to do this, but this is often done, it is a kind of heuristic approach in research. And then so, when we estimate a variogram in the 45-degree direction, which is what we will learn how to do today. What it does is, it actually considers all point pairs within a sort of conical or triangular region around the 45 direction.

So, these two arrows, all of these point pairs that are in the ambit of these two arrows will be considered. So, it only considers variation in a certain direction, when calculating the variogram cloud. So, remember when it calculates the variogram cloud, it takes pairs of points. So, in an omnidirectional variogram, it will take pairs of points in every direction.

Lecture - 51

Whereas, in a directional variogram, it will take pairs of points only along this direction lying within this conical or triangular region.

So, we will look for variation selectively and not omnidirectional. I hope that is clear and if it is not, maybe you should review the video or pause it here and think about it for a while.

Because now we are going to go to our coding session with R. So, before we do that, it would be nice to just have this firmly planted in your mind.

So, let us move to the code loading libraries, and then make our spatial data set using the data from meuse, I am going to clear all of these windows just so that we are working from scratch; let us not have anything.

So, let us load our libraries, then make our spatial data set. So, I get an error, see this is why it is important to, say could not find function make spatial data. So, the reason it could not find the function is because I had cleared my environment and when you clear your environment, everything that was previously loaded is lost. So, I need to go and reload this function.

So, remember this function was defined last time. So, it will be available in the file that we used last time, the file corresponding to session 4 and I have to source this file and when I source this file, I get this function back, make spatial data frame, make spatial data meuse. So, now I can go ahead and make this data, and call this function.

So, let us make R spatial data and then; now look at this line of code, the one that I just ran. So, this is the variogram function, you know this already, this is the log of zinc assuming a constant mean. So, typically if we do directional variograms, then we assume a constant mean. If you do a directional variogram, plus a trend removal then the results like you have to think carefully about the results because it is not obvious as to what that really means.

So, we will assume a constant mean, because see now we are estimating along the 45-degree direction, where the variation is very little. So, the constant mean assumption is actually not so outlandish. And then, we will provide our data and then we provide this additional argument called alpha and alpha is the angle, alpha is this angle, and right here is alpha.

So, 0 to 45 is 45, and then 0 to 90, so we want four values for the angle 0, 45, 90, and 135 and that is what we mean by alpha. So, we give these values here 0, 45, 90, and 135, those are the

directions in which we want to estimate the variogram and once it does that, we would like to plot the variograms.

So, now as is expected instead of one plot we get four plots and each of these corresponds to a directional variogram in a particular direction. So, let us read this plot here, this bottom right plot is the variogram that corresponds to the north, the 0 direction then 45, and then 90 up here, and then 135.

So, let us zoom this out. So, you can see it a little better. So, this is the North, North-South direction this is 45 degrees, this is 90 which is East-West, this is 135. So, one thing that we should notice right away is that the variance, the value of the semi variant which is gamma h on the y-axis is the least for the 45-degree direction. Like this variogram, it is kind of the most gentle, shallowest, and also the lowest like the actual y values are lower than all the other three.

And that is to be expected, because we just saw in the data that the 45-degree direction is the direction in which the variation in the data is the least, if you move in this direction you basically encounter very similar values. And that is why we see the least variation in this variogram.

That may or may not be a good thing depending on what you are trying to do, but we should be able to correlate this with the data. Now, there are some additional parameters that we can pass to a variogram function and these two parameters are called cut-off and width and we must be familiar with these parameters. Because you need to provide correct values for these to get sort of a reasonable result. And cut off is the maximum distance within which we want to consider point pairs.

So, think about this, you have this data. Now, when we compute a directional variogram or a non-omnidirectional variogram, it takes pairs of points. So, these two points, these two. So, every single pair of points, and computes the semi-variance between those two points, the values of those two points.

But there may not be a situation where we want to consider points, maybe that is here and all the way here, like maybe there is some cut-off beyond which we do not want to consider point pairs. So, if we are looking at this point, then maybe we only want to go to this

distance, and beyond that, we do not want to pair anything with this point, because we do not believe that there is any relationship between points that are that far away.

So, for example, if we know that zinc concentrations in a river bed are only correlated within a distance of let us say 50 meters, and you know where the river starts somewhere in the mountains and ends somewhere in the plains. Then zinc concentrations out in the mountains, do not necessarily reflect any relationship with the zinc concentration where the river ends in the delta.

So, we only want some local correlation. So, we should provide some cut-off distance. So, it will not consider point pairs beyond that distance and the cut-off value that I have given here is 1000, I just looked at this graph it goes from 0 to 1500, but I am saying that beyond the 1000 point, I do not want to consider the variation and then the width is the distance, the size of the bin.

So, how? So, if I say a 100, then basically all point pairs that are within 0 to 100 units of distance apart will be considered 1 bin and then it will plot 1 point for that whole bin, right? So, if we believe that there is a lot of small-scale variation. So, essentially the width is the distance between two consecutive points in a variogram. So, like this distance in the x direction between two consecutive points any two points is the width of the variogram.

So, what does this mean? Like if one should we have a large value or these are all semi-artistic decisions because this will depend on you there is no right or wrong answer. But it should correspond with what you know about the data-generating process. So, if you think that there is a lot of variation at small distances, then you need to have a smaller width to capture that variation, but if there is not a lot of variation at small distances and use a small width then you will capture a lot of noise.

But if you use a width, that is very large, then you will smooth over a lot of localized variation which is informative and useful. So, this is also something that you have to try and sort of play with and see at what value you get results that you consider reasonable. So, I have used a width of 100 here.

So, let us run this and then plot the variogram. So, you can see that the distance between the points increases fewer points because I am binning a lot of points together into 1. So, in the

variogram cloud, any point pairs that are between 0 to 100 units of distance apart, will be averaged to 1 like 1 value.

Let us try a width of 50 and let us plot the variogram again. So, you see now the distance between points decreases. So, there are more points now. So, we are capturing finer scale variations and also know that these are not exact. So, when you do a width of 50, if I look at this. So, my object is called l log zinc dot variogram dot directional. So, lzn dot vgm dot dir.

So, let me open that up and see what a variogram object actually looks like, the first column is called np which stands for the number of points; that means, the number of points that belong to that particular bin.

So, wait, let me show you before I do this, let me show you a variogram cloud, then it will become much more clear.

So, let's see the cloud, let us plot this cloud. So, this is the cloud right and this goes from 0 to some value just over 1500 and this is the value for every single point pair. So, to go from this variogram cloud to the variogram, I need to group these points together. So, essentially, I need to decide what will be the size of my bin. So, if my bin is 0 to 100, then I will go from 0 to about 100 and then group all of the points that are separated between 0 to 100 units of distance apart, and then 100 to 200, 200, 300, 400, 500, right?

So, it is basically how I am averaging or how I am grouping these points in the cloud to go from the cloud, if this is the cloud then what is the let me show you then the variogram to this variogram?

So, it groups the points and averages them because the variogram is actually the expected value of all of the points in a particular bin and you get this variogram. So, if I had a smaller value for width, you would see many points between these two points as well, but this is not exact. So, for example, I did a variogram here in which the width was 50 and then if I go here, it tells me the distance, the value of h for every bin. So, 0 to 82.7 or so is in 1 bin, then 82 to 132 is in another bin, and 132 to 178.

So, these values are actually not equal, they are not equal, but they are roughly equal. So, when you say a width of 50, it is not going to give you a width of exactly 50. So, I did a

width of 50 here and if I subtract the value of the distance for the second bin minus the first bin then I do not necessarily get a value of 50, I get 49.57. So, that is close, but not exact.

Similarly, if I do a width of 200 then like the distance, the sort of the difference between the 1st and the 2nd bin is 164, 197 between the 3rd and 2nd, and then 197 between the 5th and 4th. So, it is not exactly 200. So, this is just to show you that when you provide a value for width, it does not always give you points that are exactly that distance apart, but roughly that distance apart.

And then I plotted here the variogram when the width is 200. So, you see now you have very few points. So, this is typical, I think for this particular variogram I am pretty sure that this width is too broad, we are smoothing over, and we are losing a lot of local variation. And if you lose local variation then your variogram estimation can actually be typically wrong.

So, what is the right width and the right cut-off? This is something that you will have to consider carefully depending on what your research question is.

So, that is all I had for that. So, what did we do today we estimated an isotropic variogram, something that can be very useful and an isotropy is a property of most spatial processes, there are very few spatial processes that are truly omnidirectional or isotropic or the same in every direction.

And we also considered the effect of cut-off and width parameters on the final estimated result, this is just the beginning of geostatistical modelling with the gstat package, if you want more please go and read the latest manual, I do not have the scope nor the time to go into everything right now. My goal is just to get you started, but definitely check out the latest manual for a lot more ideas on what you can do, that is it for now.

Thank you very much for your attention.