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

Lecture - 50
Session - 04
Computing Experimental Variograms – Part I

Hi, welcome to another practical session. We are learning Spatial Statistics and Spatial Econometrics with R. My name is Saif Ali and today's topic is Experimental Variograms; how can we compute experimental variograms using R. We are going to be spending two sessions on this, today is the first part of those two sessions.

Before we get started, let us reiterate a mantra that all of you should be very familiar with, we are focusing in these practical sessions on developing skills; particularly developing programming skills with R. How can we do spatial statistics in R? How do we get the right programming skill for that? And, the way to obtain skill is to apply understanding to real-world problems. And therefore, obtaining skills is conditional on obtaining understanding.

So, if you do not understand the theoretical concepts, if you do not understand the principles, it will be very hard for you to obtain the correct programming skill. The two go hand in hand. And, while understanding is gained by listening, reading, absorbing a lot of material, thinking, writing. You also have to question the material that you absorb and have to solve things on your own. Solving things on your own is very important for attaining a conceptual understanding of a subject.

Similarly, when we are talking about skills, you are not going to gain skills simply by listening to this lecture. You can listen to this lecture, you can listen to 10 more of these lectures. You can read many programming books, but you will not obtain any skill; you might obtain a lot of information, but not any skill.

So, please as you see me go through the code, pause the video, go back to your own R installation which you should have by now, and try the code on your own, try changing the code, and play with the code. And, sure enough, you will have errors, there will be failures,

but that is the way to learn. Write a lot of code, keep trying, keep failing, and try again until you get the right answers.

With that said, let us talk about this particular session. What should we understand? What should our understanding be to understand the material that we will cover regarding R and the skills using R, that we will learn in this session? What conceptual understanding should you already have? Well, we are going to be estimating variograms and for that, you need to have a firm clear understanding of the idea and definition of spatial stationarity.

Spatial stationarity is a condition, it is a pre-condition, it is a decision that we make before we start applying geostatistical modeling to any data. So, this is something you need to understand ahead of time. If you do not remember or understand spatial stationarity very well, I do recommend that you pause this video and go back to the appropriate video in the theory sessions and make sure that you understand that.

Of course, you need to know the definition of a semi-variogram, what is it? It is an instrument to measure spatial variation at different scales. It's good if you remember or are aware of the estimator that is used to compute experimental variograms and also if you know what a variogram cloud is. If you do not know what a variogram cloud is that is ok, I will show you what it is.

But, you should have these ideas firmly in your mind. So, what kind of skills or what kind of things should we already have done in terms of programming? Well, remember that we made a spatial point data frame using the Meuse data set. This is something that you already should know how to do, you should know how to prepare spatial data for spatial analysis.

If you do not know that please review the earlier sessions. What will we do now? We will take a moment to look at our data and we will try to think whether spatial stationarity is a reasonable decision to make, whether should we use a specially stationary model to model our data and what kind of things should we be thinking about. And, then we will go through the process of actually estimating our variogram. We will also talk briefly about trends and how to remove them from the data and why we should remove them.

Now, as I said before whenever you want to estimate a variogram that is our goal today, we want to estimate a semi-variogram from some sample data, the Meuse data set. We have a sampling of zinc concentrations along a river bed and we want to model the spatial variation

in these zinc concentrations using an instrument called a variogram and we want to estimate the variogram from the data. But, before we start doing that, we have to know or we have to reason or argue whether stationarity is a reasonable thing to assume.

Now, remember stationarity is a property of the model, it is not a property of the data. Let me repeat that stationarity is a property of the model. What we are saying is that if we look at this realization of data on the left with the green dots, I hope you remember this plot; can we decide as an analyst to use a spatially stationary random function to model this data, to model these zinc concentrations?

Is it reasonable to say that given this realization of data, the process that generates zinc concentrations in this region is a spatially stationary process, does it have a constant mean? We cannot verify this from the data, there is no way. So, this is also something you should understand from your theory sessions, there is no way to verify whether our decision to use a spatially stationary random function is right or not. But, we can rule it out, we can look at the data and we can say no it is not right.

We cannot say whether it is right, but we can say that it is not right. So, if you look carefully at this data, there seems to be a trend. A trend is some obvious noticeable difference in values across space. This is a spatial trend. So, I have drawn an arrow here to show you that there seems to be a trend in this direction. This is the north, west to south, east direction right. Why?

Well, because if you look at the values along this outer belt, we have some big sort of bigger circles, relatively bigger circles and as you move in this direction the circles seem to get smaller. So, it seems that values of zinc concentration are decreasing as we move away from this outer belt. So, this is called a spatial trend, and for reasons that should you should know from your theory sessions trends imply non-stationarity.

If a process shows realizations that have some trend, stationarity is not a well-founded decision to make. However, we can still model this data, but we have to remove the trend. So, we have to take the trend out of it, subtract it from the values, then estimate the variogram, and then add it back. So, I will show you how to do that in a simple way, very simplified way today. I hope this is clear to you because I am going to move now to our R programming session.

If anything is not clear, please go back and review what you can. Now, is a good time to pause the video and review all of the ideas that we have discussed so far. If everything is ok, let us move forward.

And, go to our R code. This should start looking very familiar to you, as usual first we are going to start by loading our libraries.

And, today I want you to see something that we have done for the first time. I have taken some code and put it inside a function. A function in R is some code that you usually call again and again, some code that you need to run many many times. So, what you do is you encapsulate it inside a function and then you can just call that function to run all of that code for you. You do not have to repeat that code again and again.

So, what this function does? As the name suggests, this function is called make spatial data frame, Meuse makes spatial data muse. So, what that does is it takes the Meuse River data, turns it into a spatial data frame, and then returns that spatial data frame. This is something we have to do again and again. So, I have gone ahead and put it inside a function. So, the function definition ends here with this ending curly bracket. It begins with this curly bracket and this is the whole function. This is code that you have already seen.

So, I will not go into it. Just know that what it does is it makes a spatial data frame using the Meuse data and returns that frame so, we can use it. So, here is an example of the function call. What I am doing is I am calling that function and function calls always end with parentheses and I am assigning the return value to a new variable called sp data which stands for spatial data input.

So, this is the spatial data that goes into our variogram modeling exercise. And so, if we run that and then just plot the data, we get back our plot that is becoming very very familiar to us.

Now, that we have this data in spatial data format, what we want to do is we want to tell gstat, the package gstat to estimate a variogram for us and that is done by calling the function variogram; obviously, very easy to remember the name. And, if you want to look at this function in detail, you can read the manual or pull up the help by typing question mark variogram. And, then R will show you the help with a description of all of the parameters and values and other details and examples of how to use this function.

There is a lot more to this function than what I will use today. So, the first parameter of this function is a formula. Remember, this tilde operator is a formula operator and this formula has a left-hand side which we are saying is the log of zinc. So, the word zinc, remember zinc is a column, it is a variable inside our data. And, the way we know that is we can look at our data, a preview of our data by using the head command and then it will tell us that one of the columns is called zinc.

And, this column we know from earlier sessions is a column that provides zinc concentrations. This is the value of zinc concentrations at different locations on the river bed and the locations are here inside the coordinates column. So, here we are saying that we want to estimate the variogram.

But, we are not estimating the variogram of the zinc concentration directly where we want to estimate the variogram of the logarithm, the log of zinc. And, there are reasons why we do this, but I am not going to go into them right now. Just know that what this formula is saying is that we want to estimate the variation not in the zinc concentration values themselves but in the logs of those values. And, the one on the right-hand side is saying that right now we are not going, we are assuming a constant mean.

So, this one is basically our stationarity assumption. We are assuming that the mean of the random function that generates zinc concentration values is constant across the whole region. So, this one is very very important. It is just one character in an R, R code you know you can just run this code you will get some results. But, if you do not know what it means your results can very easily turn out to be garbage.

So, then the next parameter is our data. This is where we supply the zinc concentrations and the coordinates which are needed to compute the variogram and the third is cloud equals TRUE. So, what we are doing is right now we are not estimating the variogram, we want to first estimate the variogram cloud. Remember, the first step in variogram estimation is to estimate the cloud. So, let us see what the cloud is, I will explain this to you in a bit. So, let us run this, and then let us plot this cloud.

So, this is what we get. This is your variogram cloud of zinc concentrations and on the y-axis, we have semi variance and on the x-axis, we have distance. So, if you remember a variogram is denoted by gamma of h. So, it is a function of distance, it is a function of spatial separation

between point pairs. And, the value that it gives is computed using an estimator that looks like a variance estimator.

And the value that comes out of that estimator is called semi-variance. So, what this variogram cloud is showing us is that it is showing the semi-variance for every single point pair in our data, which is not very helpful. So, what we can do is we can start identifying some points. So, we can use this cross cursor and we can click on certain points and then we can click finish.

And, then what it does is it shows us the point pairs that correspond to those points on the variogram cloud.

So, those points the ones I clicked on are these point pairs. So, we can actually identify the point pairs for the variogram cloud using this interface. So, I clicked on three points and these are the point pairs. So, they are separated by some distance. So, what the variogram cloud is basically for every pair of points in this data set, it computes a value using the estimator that you already know and that is what the variogram cloud is.

So, the way to get this option to identify point pairs is when you plot the cloud, you have to use this function and you have to provide this parameter identify equals TRUE. So, this allows you to identify the point pairs. If you do not do this, then it would not allow you to do that.

Well, very good, so we have estimated the variogram cloud and we showed, we have seen it we have identified some point pairs; we understand what it is doing. But that is not what we want, we want to estimate the actual variogram. So, let us go ahead and do that.

If you want to estimate the variogram it is the same function call, but this time we do not give this cloud equals TRUE. We do not include that, we just that the default value for the cloud flag is false so, we leave it as it is. We do not provide any value and if we run this, then we get a variogram and the variogram is stored in this variable logarithm of zinc lzn dot variogram.

And of course, we want to see what that variogram looks like and this is what it looks like. Again, on the y-axis, we have gamma h which is semi variance and on the x-axis, we have

spatial lag which is denoted by h. This notation should be very familiar to you, if it is not you can easily go back to the videos and review.

And, we see that we get a set of points and these points tell you the average value of semi-variance at particular spatial lags. So, for example, at a spatial lag of 500, the semi-variance is something between 0.4 and 0.6. And, we see a pattern where for smaller values of spatial lag, we have lower variance, and variance increases as spatial lag increases.

So, what that means, is that this is evidence that our data has some spatial autocorrelation. In the sense that, values that are nearby are less variant or less at variance with each other. If the semi-variance of values that are nearby that are separated by small spatial lags is more similar and as you go further away, you encounter more and more dissimilar or more variant values.

So, now, what do we want to do? Well, remember we did not fully believe that our data is coming from a stationary random function. So, we want to remove the trend. So, how we remove the trend is we estimate the trend using some variables, using linear regression. So, we provide a formula here. So, focus on this formula. Now, this time we are estimating the variogram, but we want gstat to remove the trend before it estimates the variogram.

So, the way we do that is instead of a one, we provide some variables here, that we think drive the trend. So, here we are giving x plus y. So, this is like a regression formula. We believe that the reason we do this is that we believe that the trend in zinc concentrations is driven by location, by the spatial coordinates; remember x and y are the spatial coordinates.

If you look at the data, if you look at meuse, the data then x and y are the spatial coordinates. So, we are modeling the trend using the spatial coordinates. So, what it will do is it will predict a value for the logarithm of zinc using spatial coordinates as aggressors, subtract that value and then run the variogram estimation only on the residuals.

And, the residuals once we have taken out the trend, we can make a stronger argument that the residuals are actually spatially stationary. I mean they can be modeled with a spatially stationary random function.

This is not the only way to remove trends. Trends are usually driven by a set of complex factors, zinc concentrations can be driven by a number of factors that you will know if you have domain knowledge of how heavy metal concentrations vary inside rivers and water

bodies. This is something that experts know. And then so it depends on how well you know the domain, which will determine how well you model the trend.

This is actually I think a pretty bad model, modeling the trend is with the location. There are certainly other factors that drive the trend. So, here this is a kind of art, you know there is no right or wrong answer. You have to provide a strong model for the trend and be able to argue for that model using your prior knowledge.

For now, let us consider the trend is purely driven by x and y and estimate a variogram that is detrended and then plot that radiogram. So, we get another graph that has a similar structure. But, what we want to do is we want to actually compare whether it made any difference, does it make a difference if you control for the trend or not.

So, here I will introduce you to another library. We want to plot both of these variograms on the same graph. So, we need to make a slightly more complex graph. So, I am going to use this graphing library called Plotly which I have already installed and I load the library. And, I will not explain this code, I will let you look at it on your own.

And, if you run this, it gives you both of the variograms on the same graph and it kind of gives you a neat feature. You can look at each point and hover over it and actually see the value. And, the orange series is the variogram with the trend and the green one is the variogram detrended. So, you can see that they have maybe a similar shape, but the detrended variogram is less steep. So, the variation kind of increases at a slower rate maybe, and the absolute value of variation is smaller at the same spatial lag.

So, there is less variation in the residuals than there was in the original values. Now, again here this is sort of an artful artistic aspect of geostatistical modeling. You have to decide based on the application and what you are trying to do and what your research question is and what your knowledge of the process is, to decide which one of these will you actually use. It could go either way, I am not an expert on heavy metal concentrations I do not know.

But, this is just to show you that your variogram modeling has to accord with the theory of spatial stationarity and it actually has to serve the purpose of your research question. So, I am going to stop here and go back to an exercise.

And, I just for this exercise that both you and I will do together. Please look at the code in the first line. This is a code from the code that we just discussed and ran inside R. This is estimating a variogram of log of zinc over a detrended surface where the trend is being modeled by the x and y coordinates.

So, I want to ask you what does the formula mean? What does it mean to log zinc over x plus y? And, I have given you the answer here. You can pause the video to take a moment to think about it and come back. I will give you the answer right away. It means that we want to estimate an experimental variogram of the log of zinc after removing the trend from the zinc variable.

And, we want to model the trend using x and y as regressors. And, the reason we do that is that we believe the trend is purely a function of location, but this may or may not be true. And, it depends on how knowledgeable we are about heavy metal pollution or heavy metal concentrations. So, that is the answer to this exercise. Please review it, if you have not understood.

So, what did we do today? We wrote our very first R function. We estimated a variogram from the Meuse data set and we compared variograms with and without trend removal. And, we learned about a new library plotly and we learned some new functions and that is it for now. I will see you next time.

Thank you very much for your attention.