Lecture - 52

# Practical
# Spatial Statistics and Spatial Econometrics
# With R

Prof. Saif Ali
Department of Social Sciences and Humanities
Indraprastha Institute of Information Technology, Delhi

## Session - 06
## Lecture - 52
## Fitting Model Variograms in R

Hi, welcome to another practical session in which we are learning Spatial Statistics and Spatial Econometrics using the R programming language. My name is Saif Ali, today's topic is Fitting Model Variograms and this continues from previous topics and previous understanding. We are focusing in these sessions on developing programming skills using R that are needed to conduct Spatial Statistics and Spatial Econometric operations.

But the acquisition of these programming skills relies critically on conceptual understanding which you will obtain by watching and learning from the theoretical lectures. So, for each practical session, we go through the conceptual understanding and the theories and principles that you need to already know to maximize your benefit and maximize your skill acquisition from these settings.

Today in these sessions, what would be great is, if you for your conceptual understanding, what we should already know, is the idea of variograms both omnidirectional and directional variograms, which is something that we have done in previous sessions. You should know what they are, you should also know how to estimate them using R because this is all material that is already been covered.

It would be good to have a notion, and some understanding of model fitting, how do you fit a function to some data? What does it mean to fit a function to data? It would be nice if you at least conceptually know what this is; what is the meaning of fitting a function to data, this is something that we will do today, but I will not be able to go into the theoretical background.

So, if you are familiar with model fitting good, if you are not I recommend you pause this video and review some material online or from some textbooks or maybe ask your teachers and friends.

**Lecture - 52**

This is something that you will need to understand to appreciate the material fully today, you should already know from this course variogram models. The spherical model variogram, exponential, and others what they look like and how do they differ from each other and what are the parameters of a fitted variogram, these terms, the sill, the nugget, and the range these were covered in this course in the theoretical sessions.

So, please go and review them if they sound unfamiliar and if they are familiar to you then that is great. So, in terms of programming skills, what should we already have done? Well, we should already know how to estimate variograms using the meuse data set, both directional and omnidirectional variograms, this is something that ideally you should have gone back and done many times on your own and played with the various function arguments and parameters.

So, it is something that you really, it is kind of you know from the back of your hand. So, we can build upon it and so, today what we will do is we will change things up a little bit and we will use a different data set. We will not rely on the meuse data set for heavy metal concentrations that we have been using, we will import some data that we get, this is real-world data that we got from a government data portal and then we will fit a model variogram.

So, we will estimate that we will load a new data set, we will estimate a variogram, and then fit a model variogram to it, that is what we will do today.

It is just a quick review of what is an experimental variogram. Well, it is a function of spatial lag as you see in this figure, on the x-axis we have spatial lag denoted by h and this gives us a value of variation or variance for some set of spatial lags for some discrete points on the x-axis.

So, for h values 1, 2, and 3 maybe 10 to 20 values of spatial lag, we know the value of semi-variance. So, we have some idea that these different values of h, this is the value of variance in zinc concentrations for example, but that is not what we want, we actually want what we want to do in the future, we want a value for any value of h we want to know the variance at any value of spatial lag.

So, even for example, from this experimental variogram, we do not know how the variance varies between this point and this point, all the points that are in between, all the values of h, we do not know the behavior of the function That is important for us for what we want to do in the future. So, what we can do is one idea is that it is that we use in geostatistics, we try to fit some smooth function through these points and the idea is that this smooth function is continuous.

So, we can get a value of variance or semi-variance for any value of h without sacrificing a lot of accuracy because we believe that this function fits these points pretty well. And we should also be able to say or quantify how good the fit is, if it is not a very good fit, then we have to question our choice of function and try a different one and we should also be able to compare one fit from the other. So, if we fit two different functions we should know which one we prefer based on some quantified measure.

That is the basic idea of fitting a model variogram, this red curve that I have drawn, I have just drawn this by hand, but we want to use a closed form function and this is called a model variogram. So, the initially estimated points, those set of points is called the experimental or the estimated variogram, and the fitted variogram is called the fitted model variogram or just simply the model variogram.

So, this is what we want to achieve today using R on a brand-new data set. So, if this is clear, that is great we will move forward if anything is not clear please go back and review the video or previous materials if you need to you can pause the video now and go back and I will be right here when you come back.

Meanwhile, I will move on to our live coding session with R.

So, let us go to the code, we have some new stuff today. I will go through it one by one. So, this stuff you know lets us load our libraries gstat and sp.

So, the first thing we want to do is, we want to load a new data set we are not working anymore with the meuse data set, we will work with groundwater level data. So, you know the groundwater level is the depth at which the water table exists beneath the surface of the ground.

So, if the surface of the ground is at level 0 and the water table where you start getting water from the underground aquifer is at let us say 3 or 4 meters, then that exact measure of where the water table is from the ground surface is called the groundwater level. So, in that case, it would be 3 or 4 meters.

And the government of India and various states in India monitor groundwater levels all over the country using observation wells and this data is available publicly on the India Water Resource Information System WRIS. So, I encourage you to check out this portal, I have already downloaded and formatted in sort of clean this data for you, and we will make it available for you.

So, I have it here on my computer inside a folder called data as a CSV file.

So, remember CSV means Comma Separated Value files and this is basically just tabular data. So, the way to read CSV files is using a function called read dot CSV and we stored this object in this R object called up dot gwl which stands for Uttar Pradesh groundwater level.

So, let us read this file and now we want to know what is in this file, we do not know what we have loaded. So, we first want to look at it.

So, let us open it up by clicking the object here and looking at the column names. So, we have a district that is simple enough, this is the district in which the groundwater level observation is being made and the block which is a sub-district administrative unit, and then the hydrograph station which is the name of the station, where the groundwater is being monitored.

This is the longitude and latitude of the spatial coordinates of that station of the groundwater monitoring station as you can see we have spatial coordinates for some of the stations, but not all, some of them are missing and we will have to remove these missing values. Because if you do not know the spatial location of an observation then you cannot really use it in spatial statistics so, we must know the spatial location if the long latitude and longitude missing there is not much we can do. So, we have to remove those observations.

Type of station, I do not worry about that too much, here this is the year in which the observation was made and then pre-monsoon and post-monsoon. So, in UP they make two observations one in pre-monsoon, which is around May just before the rain start, and one in post-monsoon in September when it ends. Or when most of the monsoon ends the idea is that they want to get a sense of how much the rain recharges the aquifer.

So, they measure the level before the rain and then after the rain and then they compare how much the level rows are because of the rain. So, that gives you a sense of how much water you have to spend for the rest of the year. So, this is what the data looks like, of course, we could have also just done head up remember head, the command head. So, if you just type head it would have shown you the same information, but I preferred you just preview the data by clicking on the name of the objects here.

So, now we know what is inside, and what we can do. So, we are not going to estimate the variogram for all of the data. So, as you can see there is a column called year and so, this data is for many years of monitoring. So, for every year at every well they made two observations and then this is so, the data that is been collated.

So, we can use the function called range to check how many years we have in the data and the range is from 2009 to 2018 which is the minimum and maximum value for the year.

So, let us say we want to estimate the variogram for a particular year, somewhere in the middle of that period. So, I have chosen the year 2015 and this is the new command that you do not know yet. So, this is called a subset. So, what this does is, from this data object it picks only the observations where the year column equals 2015.

So, this will just pick out the rows for the year 2015 and store them in a new object, which I have indicated by dot 2015. So, UP groundwater level for 2015, if you run this, we will get a new object. So, we will select data only for the year 2015. So, if you see how many rows there were in the first, there were about 88,490 rows, but if you look at how many rows there are in your in for the year 2015, it's only 193. So, we only have 193 points, sorry that is not what I meant to do, it is not Western UP yet, it is UP.

So, we have 8849. So, it is considerably less, obviously because this is the data just for 1 year. Now we want to go further. So remember this, you know this is something new subset, this is something we have not encountered. So, we can use a subset again, for this exercise I do not want to estimate the variogram for the whole state of Uttar Pradesh.

I just want to do it for some districts in western UP and the reason for this is of course, efficiency because some of the routines to estimate these things take a while and if you do it for the whole state we will be here for a long time. And also sometimes doing things for the whole state is stationarity is a problem because you know you cannot argue that there is one process that governs groundwater level dynamics in the whole state.

But for Western UP you can make a stronger argument, you can say it is kind of under one river basin, one aquifer conditions are more uniform it is a smaller area. So, stationarity is a more reasonable sort of assumption. So, again I subset my data for 2015 and my condition for sub setting is that this district column.

So, this is the new operator called in. So, this is the word in sandwiched between 2 percentage signs what this means is that the district should be one of these, it should be in one of these values. So, I only want data for Muzaffarnagar, Ghaziabad, Baghpat, Meerut, and Hapur.

So, let us go ahead and run that. So, now, we want to see how many observations we have in this. So, we had 8000 or so, observations for 2015, but if you reduce the spatial extent just to these districts we are left with 245 points. So, that is kind of what we are working with about 245 observations.

Now remember, we have to convert this into a spatial data frame where we cannot just use it directly. So, let us convert it to a spatial data frame, now I get an error here. So, it says missing values in the object right? So, I told you that you cannot convert something to a spatial data frame if it has missing coordinates, it does not know what to do where the coordinates are missing. So, we have to run this line of code which is called na dot omit this is the function that removes all observations where there are missing values.

So, in this data, if there are missing values in any of these columns it will remove the whole observation, this is something called listwise deletion. So, we have to do this now; obviously, we are losing data this way. So, this is not something you should do very sort of without thinking about it, this is something you should document, but for now, we are going to remove these because we simply cannot do anything for observations where we do not know the spatial coordinates.

So, let us get rid of the missing data and then run the coordinates. So, now, this time it did not complain.

**Lecture - 52**

Now we have a spatial point’s data frame, let us make sure we have a spatial point’s data frame yes so, the class or the type of our object. So, our data is now contained in this object called western up gwl dot 2015. So, these are the groundwater levels for western up for the year 2015. Now before we go forward we are going to do one more thing, we want a map of up because when we know we are going to plot we also want to see our values on an actual map.

And maps are available in something called shape files and to work with shape files you need this library called rgdal. So, you will have to install this I have already installed it. So, let us go ahead and load it and then read a shape file for India which is read by a function called read ogr, and I put the shape file in the data folder and we will share this with you.

So, let us read this shape file and then again we do not want the whole of India.

So, if you plot this shape file we will get a plot of all of India, but that is not what we want we only want the districts that we care about for districts. So, we select those districts from the shape file and now if we plot it again we only have those districts.

So, that is the kind of study area, we want to study groundwater levels in these four districts for the year 2015.

So, before we estimate anything, let us just plot our data in a nice plot I am not going to go through this code in the interest of time, I encourage you to read this code on your own. This is not code that is doing anything related to spatial statistics, it is just a code to plot a set of points and a map together on one graph, this is done with a library called tmap. And if you go through this code, you will figure out what is happening, I am just going to run it and I am going to show you what the map looks like. So, it returns an object called west up dot map, and then if we plot that we see our data plotted on a map.

So, let me zoom this in for you. Oops, our legend entry is wrong, it is as percent missingness. So, we do not want that we want to name it what it actually is, which is a groundwater level and these are post-monsoon groundwater levels. I have selected the post-monsoon column. Remember we had a pre-monsoon and a post-monsoon. So, I am going to estimate a variogram for the post-monsoon observations.

**Lecture - 52**

So, groundwater level post-monsoon, Western UP 2015, is actually just a part of Western UP, but that is good enough for now as a title.

So, let us run this again and let us see the map again and now we have the right title groundwater level post monsoon Western UP, let us zoom this in and see.

So, remember the first thing that we did last time was try to spot if there is a kind of trend, is there some obvious trend, if there is an obvious trend then stationarity is not a valid assumption.

I am going to say, I cannot really see a trend in this data, but it does seem like in this area the circles are a bit bigger whereas, here they are a bit smaller, then here again maybe a little bit bigger, but I do not see an obvious sort of direction in which the circles get progressively bigger to smaller or smaller to bigger.

Some other people may argue otherwise that is why a lot of these things involve an artistic kind of modeling decision, but for now, I am going to go with a stationarity assumption and assume a constant mean. So, remember that when we assume a constant mean, we estimate a variogram using the formula where the right-hand side is 1, basically we are saying that the mean is constant, we are progressing on a constant, and we are not estimating the trend using some covariance.

So, let us estimate our variogram and then plot it, and we get something that looks like this.

This is our experimental variogram. So, what we have done is we have replicated the same process from last time that we did on the meuse data set, this time we have done it on a new data set. So, we have not done anything new, we have just got some new data you should know what this is hopefully from your previous knowledge. Now, we want to go further, what we want to do today, is that we want to fit a model variogram to this experimental variogram and the way we do that is by using a function called fit dot variogram.

And this function as the first parameter takes the estimated variogram, what are we fitting to and as a second function it takes something called a model, which is given by creating an object of type vgm, and if you just hover over this.

**Lecture - 52**

The first parameter of vgm is something called a partial sill, the second is the model, the third is the range, the fourth is the nugget, and then a bunch of other parameters. So, if you look at this kind of variogram.

So, that is weird because a partial sill, nugget, these are things that we find out after we have a fit. So, why is it asking us to provide them before we fitted anything? Well, this is where we do not put the exact values, we put our best guesses. So, remember a partial sill from your conceptual theoretical lectures, a sill is the maximum value of variance.

So, here the maximum value of variance is something like 50 or somewhere between 40 and 50. So, as a best guess, I have put 40, this is the model. So, what function do you want to fit to your variogram? So, here I am fitting a spherical function.

So, the code for that is sph. I am also doing another fit where I am using the exponential model. So, I am going to try and compare these two fits, 0.4 is the range like. So, the range is the lag the distance at which it reaches the maximum values. So, I am going to say the variogram kind of reaches a maximum value somewhere around here.

So, best guess 0.4 and then the nugget. So, the nugget is to remember the variance at h equals 0. So, if I just extrapolate, so, maybe the nugget is somewhere here. So, as a best guess I have given 5. So, what is expected is that we look at the estimated variogram and we make our best estimates, and pass them here and then it will use an iterative algorithm, a weighted least squares method to try and fit the best to arrive at the best fit.

You can use different methods for fitting, you can use unweighted least squares, you can use restricted maximum likelihood and if you want to explore different methods, I encourage you to read the manual the way to do it is by using a parameter called fit dot method in which you can provide, which method you wanted to use when fitting the function, by default it uses Weighted Least Squares WLS.

So, I am just going to leave it at that, I will not go into that, I am going to fit a spherical function and an exponential function and then I will plot both the experimental variogram as well as the fitted spherical variogram on one plot and we get something like this. So, just visually looking at it seems to be a reasonably good fit, the line seems to fit the points well, but we would like a quantitative measure for goodness of fit.

**Lecture - 52**

So, one measure that sometimes people use is the sum of squared errors for the weighted least squares that were used to fit the function to the points.

And for this spherical, the weighted least squares is something like 159 0739 and then the sum of squared errors for the exponential one is 363. So, it seems the spherical function is a slightly better or rather a better fit. So, that is just a measure that you have to know, you know which one is a better fit to justify in your reports in your papers.

So, it seems for us in our case this spherical function is a better fit. So, I am going to stick with that for going forward.

I am going to stop there and summarize what we did today, we estimated a variogram using a new data set for groundwater levels, we kind of left behind the meuse data set, and we learned how to load and plot a shape file which is a useful thing to do for spatial stats and then we fitted a spherical variogram model to our estimated variogram, we came across two new libraries tmap and rgdal.

And I hope you find this session useful, please review it, please play with the code on your own, try to change the parameters, try using different models to fit your function, and once you are comfortable we will see you in the next session.

Thank you very much.