Lecture - 53

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

Session - 07
Lecture - 53
Kriging and cross-validation in R

Hi, welcome to another practical session on Spatial Statistics and Spatial Econometrics. We are working with the R programming language. My name is Saif and the topic for today is Kriging and cross-validation.

We are working on acquiring programming skills for spatial statistics, but acquiring these skills is founded on our solid understanding of the conceptual material.

And for today's session, you will need to know what a model fitting procedure is, you will have a good understanding of what it means to fit a model and what it means to fit a model variogram, which is something that we did in the previous session. You will need to remember or at least have a passing eye collection of kriging estimators and kriging equations, the process of kriging, what it means, and what does it.

And of course, you know what ordinary least squares is, which is not necessarily covered in this course, but this is something, it will be helpful for you to know because we will be using it in today's session.

For programming skill-wise, in the previous session what we did was we estimated and fitted variograms using two data sets, using the meuse data set as well as the western UP groundwater level data sets. And this is something that hopefully that we are very comfortable with because we are going to build on this today.

What will we do today? So, we will look at the parameters of our fitted variogram. So, the main parameters that tell us the shape of our fitted variogram, and are called sill, nugget, and range, I hope you know what those mean, if you do not, please refer to the appropriate video and review that material.

Lecture - 53

And then we will use our fitted variogram to perform kriging, spatial prediction with ordinary kriging. We will try to predict the values of groundwater levels at locations where we have no data from the observed data. And then we will cross-validate our model variogram selection. So, we will provide an additional way to quantify or measure how good the selected model is and how much agreement there is between the predicted values and the observed values of groundwater levels.

If that is clear, if some of that is not clear, please pause the video now and go back and review the material, if that is clear let us move ahead. So, remember this is as far as we had got last time, this set of blue points is your experimental variogram for groundwater levels. And this line is the fitted model variogram for which we used a spherical variogram model. Now, what does this give us, why does it help to have this kind of smooth line fitted to these points?

Well now, for any value of h, even between these points, we can get a value of semi-variance. So, we know for any separation, any two locations no matter what their separation is whatever the value is, we will get a value of covariance or variance between those two locations. So, we have a kind of spatial structure over the whole region, where we can select any two points arbitrarily and get a value for the variance between the groundwater levels at those two locations and no matter where they are.

So, this is a powerful kind of information to have and we can use this for spatial interpolation using the kriging estimator. I am not going to go into those equations; of course, I expect and hope that you will be familiar with them. So, let us move on to our coding session.

So, here we are, we had fitted a spherical model variogram and we had stored it in an object called lzn, still using lzn which is the log of zinc where I should be using gwl. So, please pardon me for that. The object should be correctly named to reflect what they contain.

So, lzn dot fit. So, we want the spherical fitted model. So, if we just print that out it gives us; so we can also look at it by clicking on it here.

So, it gives us some information. So, it gives us a kind of 2 by 2 matrix, and in row 1, we have something called nugget or nug for short. So, this value is the nugget effect. So, it is basically this value here, where it crosses the y-axis and the nugget effect which is the variance at spatial lag 0.

So, typically there should be no variance, but because we are estimating things from data and we fitted something to data that is not exact, we get some residual variance even at 0 spatial lag and this is called the nugget effect. And roughly, for us the value is 2, 2.2 or so, and then for the spherical model, the partial sill, this is the maximum value that the variance sort of reaches and the partial sill is at about 42.

So, the sill is where the variance increases as spatial lag increases. So, as we go further away, the variance between the observations is more, sort of at variance with each other, and that continues to increase. But at some point, it does not increase anymore. If you go further away from that point, then there is not necessarily any increase you kind of reach a global maximum variance.

And this is a very classic sort of structure that we see in spatially auto-correlated data. So, the value of this maximum variance is something like 43 so, that seems right, it's right about here. So, one thing to do is to just compare this with the sample variance. So, we can compute the sample variance of the post-monsoon and remove the missing values.

So, the sample variance is something like 41. So, this is good news because the variogram, what the variogram is telling us the value for the variance of the whole sample is the same as this global variance that the variogram reaches ultimately. So, that is a good agreement between our variogram and a different estimator for the global variance.

And then this parameter here range, this is the distance or the lag at which the maximum value is reached and this is about 0.26, which is right about here, on the x-axis 0.26. So, 0.26 units of distance, what that means is, that this is the range of spatial autocorrelation, the values of groundwater levels are correlated within this range.

So, if you have a well somewhere, then you can expect that in this radius of 0.26, the values will be similar to what the value is at this particular well or this particular location. And then beyond that, we do not expect not necessarily to be that similar. And the similarity of course varies, at 0.1 it will be very similar whereas 0.2 little less similar beyond that not necessarily similar.

So, those are the three the nugget, the sill, and the range are the three parameters that kind of tell you the shape of the variogram. And these three parameters are very important to know

and the way to know them is simply by, you can just type the name of your fitted variogram object.

So, let us for a moment see what it is for the exponential model. So, for the exponential model, you get a slightly larger sill, which is close to 50. So, we had 42, and then for the range you have 0.13, which is kind of half the range, and the nugget is 0. So, that is good, it actually gives you a nugget of 0, which is what you theoretically expect.

However, notice that the range is half of the range that is the spherical model. So, your choice of model alters the parameters and the estimates for the parameters. So, you should be careful of course, we saw that the spherical model was a better fit. Since it was a better fit, we selected that. So, we trust these values the ones for the spherical model a little more.

Now, that we have this model variogram, we want to go further. So, if I just show you the map, we have this data right we have this data, but this is not a lot of data, you know this is quite a large region, we have roughly 200 or so wells, 250 wells we want to know the groundwater level in between these wells as well.

So, what we want to do is we want to spatially predict, knowing the values at the observed locations and the structure of spatial correlation, a spatial variance that we have estimated using the variogram. We want to predict statistically the values of groundwater level at all points in this region.

So, we will need a different kind of spatial object for that, which is called a spatial grid or spatial pixels. So, we will have a grid of values and then we will estimate the values of groundwater levels, we will predict the value of groundwater level on each point in that grid. So, let us go ahead and make such a grid and that is done using this code to create an empty grid, where n is the number of cells.

So, this is 200000 cells, which is a very high-resolution grid, you do not typically need such a high resolution, but I like to have good quality predictions; or rather good resolution in my predicted output. So, I use a lot of points. So, we can go through this and this is the same thing, we need to provide the coordinates and then at the end of this, we get an object called grd which is a spatial grid object.

Lecture - 53

So, if you say class grd, it is of type spatial grid. So, it is a different kind of spatial data, so far we had spatial points and spatial points data frame. Now, you know a different kind of spatial data and then we need to project this grid to the same projection system, that we had for the wells.

So, let us go ahead and do that, and then we can use this function called krig. So, this is named after somebody called Krig, who first proposed this technique, now this function is new. So, let us go through the parameters. So, the first one is a formula, so we want to predict the post-monsoon groundwater level with again a constant mean assumption, we are assuming constant mean stationarity.

And then we want to subset our data with only the observation which are not missing. So, this is not necessary for us, because we had already removed the missing observation previously. We could have just passed in the data and then we pass in the grid that we just created in the lines above and then our fitted model variogram.

So, it needs the formula, it needs the data, the observed values which are these blue circles; the grid on which the prediction will be performed, and the fitted model variogram using which the prediction will be performed.

So, if we go ahead and run this, it is using something called ordinary kriging because we used a constant mean assumption. So, we got a warning that is fine, we can ignore that for now. Although one should typically not ignore warnings. So, it does not like the use of is, na. So, it does not like this line. So, I think if I just remove this and just pass in the data hopefully, that will just go away.

If I just remove this, yeah ok, so, that went away alright. So, now, we have a spatially predicted groundwater level data set, where we have a value for every point on a very dense grid, but the thing is we need to be able to actually see it. So, to plot this, I have added two functions map predicted groundwater level and map variance.

So, what this does is if I pass in the krigged object, the spatial grid, it will give me a map of that grid. And I am not going to go through this code; I encourage you to look at it yourself. This is using the raster library and the tmap library some of which you are already familiar with.

Lecture - 53

In the interest of time, I am going to call this function map predicted groundwater level and pass in my krigged object. Could not find the function raster, yeah. So, this is the reason, we have to load our libraries. We never loaded our libraries, right?

So, it could not find a function.

The raster function is available inside the raster library.

So, we needed that. Now, that we have done that let us try and ok. So, there you have a krigged or a spatially interpolated or a spatial prediction of groundwater level over the whole region from a set of 250 or so observed values and we see that there is a large sort of hot spot of depletion around Baghpat and then one around the urban area of Meerut and for the rest, the groundwater levels look a little better.

Now, this is a statistical prediction, done using the Kriging estimator. So, typically we expect to have a variance value for every one of these predictions, we need to know how confident we can be of this prediction.

So, what we can do is we can also map the variance, we can map the variance of each prediction and we use the same object.

And this is a map of variance. So, you can see the darker reds mean higher variance and I will show you why exactly there is darker red along the edges and slightly light. So, basically what this is saying is that we can be more confident of our predictions where there are lighter red colors and a little less confident where there is more red and not confident at all when there is you know sort of deep red shade. So, this is just a measure of how good our prediction is.

So, let us go back to this for a while and now let us look at what we have done. So, we started with this data, data that somebody went out there to an observation well lowered a tape into it, and measured the groundwater level in the year 2015 at 245 locations in post-monsoon in this western UP region.

And we took that data and applied a geostatistical model and using the variogram that we estimated from this data, we estimated the structure of spatial variation and we spatially predicted purely from a statistical point of view. We have not gotten into any hydrology, we

do not need to know what the subsurface flows were like, and we do not need to know anything about the process. I mean we do if we want to argue about stationarity.

But our prediction is done purely with geostatistical methods. We prepared this map and of course, as you can see the circles are a bit bigger here in Baghpat. So, we expect these predictions to be also larger, and then circles are a little smaller. Now, see if the variance here is very high, well that is because in Ghaziabad we do not really have a lot of data.

So, if we do not have data, we are not going to be very confident in our predictions which is what this variance is telling you. So, if your prediction variance is high, then you need to be careful about what you use the data for, and how much confidence you can place in your prediction. So, this is very useful to have and this is something that you do not have with physical spatial interpolation methods like inverse distance waiting or other methods which are deterministic and are not statistical in nature.

So, having said that, let us go back to our R code. Now, remember we had computed the sum of squared errors for our fitted model variogram to see how good the fit is. We have one more way after we have done prediction, what we can do is we can do something called cross-validation using a command called krig dot cv. And cross validation what it does is, it takes out one observed value of groundwater level and predicts it using the rest of the data.

So, then at that location, you have an observed value as well as a predicted value. So, you can compare how good your prediction is with what was observed. And this is a sort of often used in the research literature to estimate how well you have done with your model selection and model fitting.

So, let us go ahead and do that and then let us regress our predicted values on the y-axis on our observed values.

If you look at the object produced and you look at the data, we can just look at the data, it returns an object called lzn dot cv.

And we can look at only the data of that object. So, you have a predicted value in the first column, where 1 dot is predicted and then where 1 dot where this is the prediction variance. And then the observed value at that point and the residual fit when you and the z score of that residual.

So, what you can do is, you can plot the predicted value and the observed value. So, that is what we have done here, this is a scatter plot of the prediction versus the observed value.

And then you can regress this and draw a regression line and then try to get a summary.

So, you can see that the slope of this line is about 0.63 or so. So ideally, we want the prediction and the observed values to be exactly equal. But we find here that on average the predicted values are less than the observed values, they are about 63 percent of the observed values, and there is some underestimation.

We can also perform a correlation test between the predicted and the observed values and we see that the Pearson correlation coefficient is 0.77 or so.

So, the values are highly correlated, but there seems to be some underestimation. So, if you wanted a better cross-validation, if you wanted more agreement between predicted and observed values, you would have to go back and question some of your assumptions, some of your methods, some of the choices of model, and some of the parameters that you used, to achieve a better fit.

And this is all something, these are things that will be gained by experience and your understanding of geostatistical modeling.

So, just to summarize what we have covered today, we looked at variogram parameters, sill, nugget, and range. And these are the parameters that kind of parameterize or tell you or contain the shape of the variogram function. They each have a meaning and we saw visually how we can plot them as a graph in R and what each of them means.

We also performed spatial prediction using ordinary kriging. We made a constant mean assumption and we performed ordinary kriging and we use that to create a map of groundwater levels, over a region of a subregion of western Uttar Pradesh. And then using kriging we tested how good our model selection was by cross-validating our spatial prediction.

So, the way that worked is you predict one value at a time. So, you go to a place where you have observed the groundwater level, you remove that value from the data set. So, you have an observed value at that point, you predict the value at the same location using the rest of the data.

Lecture - 53

And then you compare the predicted value with the observed value and that gives you a sense of how good your model selection is. You can repeat the same process with another model and compare how good the agreement is between your predictions and your observed values. And that gives you a measure of which model to select for your particular research task. And that is what we did in this session; we will see you next time.

Thank you for your attention.