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

Lecture - 49
Session 3

Hi, welcome to yet another practical session where we are learning Spatial Statistics and Spatial Econometrics with R. Today's topic is working with Spatial data. My name is Saif Ali.

Before I get into today's topic, I would like to recap as usual that excellence and mastery in spatial statistics or any other subject are gained by a combination of understanding and skill. And while understanding is gained by listening, reading, thinking, and reflecting or writing; skill is acquired by actually doing, trying, failing, writing code, and trying again.

The wrong way to acquire a skill or the way not to acquire a skill is by just simply watching videos or reading many books, if you do that you might acquire a lot of information, but not a lot of skills. So, I encourage you as we are going through these videos to use pause and play frequently and when I say play I do not mean just play the video play with the code.

So, as I am writing code pause the video, stop me, you can stop me anytime you want using the pause button and go and play with the code. Try different versions of it, try and break it, and then once you are done, you are fairly confident, you can come back and play some more of the video.

So, you should be going back and forth, you should have two windows open –the window in which you are watching this video, your browser, or whatever, and another window where you have RStudio and you should be going back and forth and trying stuff out for yourself.

With that said, what should you already know and what are you going to learn in this session? Well, what you should know in terms of the conceptual understanding you should be familiar with the idea and concept of spatial location, especially 2D locations on a 2-dimensional XY plane, you should be familiar with coordinate systems, how to locate a point on a 2D plane and you should be familiar with the types of spatial data.

When I say the types of spatial data, I do not mean data types in R; I do not mean data frames, numerics, or literals. I mean the types of spatial data that you have learned in this course, I mean geostatistical data, lattice data, and point pattern data. Say you should know the difference between these three, if you do not know I encourage you to go back and watch the appropriate video in this very course, and learn what these three different types of spatial data, and what they mean.

In terms of skills, what should you have already done? Well, you should already have installed R; you should have installed the necessary packages – gstat and sp. They should be familiar to you by now. You should have also loaded the meuse data set which is the data that we are going to work with.

And, so far we worked with it as tabular data. So, if for us the meuse data set was just a table. It was just rows and columns that contained some data, but we want to go further. We do not want to look at it simply as tabular data anymore, we want to look at it as spatial data. So, we want to look at it as data that has a spatial location associated with each observation.

So, this is something that should be crystal clear to you after you finish today's video, is that there is a difference between just some tabular data that you have in a table versus spatial data where every observation is necessarily associated with a particular spatial location. So, remember that spatial data means that every observation is associated with some location. It is it occurred at some spatial location. So, space is inherently part of the data.

And, we are going to see how to manage that in R, how to convert regular tabular data to spatial data, and then we are going to see the difference that they are two different kinds of data types. And, then we will make a spatial plot that will show you the spatial distribution of the data which as you will see will be more informative than simply a histogram.

So, if that much is clear if you do not have any of the prerequisites that I just described, if you have not installed R, and you do not know what the packages are, please go back and review the previous videos and make sure you have gotten to this point before you go any further. So, now, would be a good time to pause, make sure that you have got all the prerequisites in place.

Alright. So, I am going to go to my RStudio in which I have some code that I have written for you which we will go through line by line as usual. So, you already know what this does. We

have to load our libraries and then load the data set called meuse and remind me what this function here does class meuse what are we doing here. You should know this by now basically we want to see what kind of variable is meuse, and what is its data type.

And, we see that meuse is a data frame and a data frame is another word for tabular data. It is data that is organized as rows and columns. Every row is a different observation, every column is a variable. Columns can be of different data types, but within the same column, all of the data has to be of the same type. This is all stuff from last time.

So, far so good. We have a tabular data set in a data frame called meuse, but that is not what we want. We want spatial data, we want to associate every observation with some spatial location. So, let us start by making a copy of our data set. So, let us make a new variable called meuse dot sp. So, the dot sp as you guessed correctly stands for spatial. So, we want to make a spatial version of the same data. So, let us start by making a copy.

And, this operator here, this arrow left arrow operator, which is the less than sign and a hyphen, and this is similar to an equals operator, it assigns the right-hand side to the left-hand side. So, what this does is it takes your meuse data frame and simply copies it to a new variable called meuse dot sp. So, now we are going to make changes to this meuse dot sp.

So, what change do we want? We want to convert meuse dot sp which is currently still a tabular variable to a spatial data format. And the way we do that is by using a command called coordinates and, coordinates is a command. If you look at the help manual for coordinates, coordinate is a command that comes from the sp package. This is the package that we have already loaded and this is the package that provides methods for spatial data.

So, in R documentation if the function name is given first and then within curly brackets it tells you the package name. So, our function is called coordinates and the package that it comes from is called sp, and you can read this description to see what it does. Basically what it does, in short, is that it assigns coordinates for every observation in the data set meuse. And, how does it assign coordinates? Well, it takes them from the data itself.

If you remember, if you look at the meuse dot sp, if you look at the meuse data set then the first two variables are called x and y. So, these are the locations at which these observations were made. So, for example, in row number 1, you have a location and then a bunch of observations for cadmium, copper, lead, and zinc concentrations and some other variables.

So, that means that these are the concentrations that were found at that particular location. So, we already have coordinates in our tabular data, but we need to tell R where to find them. So, we use a formula, this is called a formula anything that starts with this squiggly tilde operator is called a formula in R. I am not going to go into what a formula actually is.

But, just know that this is a formula and what this is telling R to do is to pick up the x and y coordinates from the x and y variables inside the data frame. So, it is going to treat these two as the spatial coordinates of these observations.

So, if you run this, well, I am getting an error because, so let us read the error, setting coordinates cannot be done on spatial objects where they have already been set well. So, the reason is that I have already done this previously.

So, what I am going to do is I am going to remove this meuse dot sp object and I am going to repeat this procedure just so you can see it. So, I am going to recreate it by copying the data from meuse.

And, then I am going to run this coordinate command here and now it has assigned coordinates using the x and y data. So, if you now look at the class of meuse dot sp it is something called a spatial points data frame. So, that is different from meuse. The class of meuse is simply a data frame, but meuse dot sp is a spatial points data frame. It contains exactly the same data, but now R has been made aware that it contains a special sort of data and in this case, it contains spatial data.

So, if you want to work with spatial data in R, you are going to have to become familiar with spatial data formats. And this particular one is called spatial points because our data is made up of points. We have the concentrations of these heavy metals at different points, each observation is a point. You can also have grids, polygons, and other geometric entities, but in this case, we have spatial points and for each point, we have some data. So, meuse dot sp is a spatial points data frame.

And, here I have shown you that you can do almost anything with meuse dot sp that you can do with meuse. So, a spatial points data frame is also a data frame. So, it is a more specific sort of data frame. So, for example, you can use the bracket operator to examine individual elements. You can look at individual columns. This is all stuff that we had done with tabular data. So, you can just do the same stuff with spatial data.

This is the zinc column, if you want to look at only the first few you can do a head command.

So, you can do summary stats for zinc, you can do all the same stuff. So, nothing much changes except that R now treats the data as spatial data. You can compute variance and standard deviation.

You can even make a histogram. So, remember this histogram from last time it tells you the frequency distribution of zinc concentrations. Now, here is what I want to bring your attention to. We want to make a spatial plot. So, the histogram tells us nothing about space. We do not know how these values are distributed in space and that is what we want to see. So, we going to use this function called bubble.

This function makes a bubble plot using the meuse spatial points data frame and it looks at the zinc column only. And, it uses some colors, we have provided some colors, and we have provided a title for our plot.

So, if I run this, we see that it creates a spatial plot of zinc concentrations. So, how do we interpret this plot? Well, each point is an observation that has been mapped to some location in space remember because now we know where it occurs in space. And, the size of the bubble and the size of the circle tells you the magnitude of zinc concentration. So, this plot in addition to telling you about the magnitudes also tells you about the spatial distribution.

So, it gives you richer information or different information than the histogram. That is all the code that I am going to look at today, before I go to my discussion please pause the video and make sure that you can get this far on your own and further maybe if you try some other things you can get further.

So, let us go back. So, just to summarize, last time we saw how to make a histogram. We did it using the hist command and it told us about the frequency distribution of how the magnitudes of zinc concentrations are distributed, but it told us nothing about their spatial distribution. By doing a little more work converting our data into a spatial data format, a spatial points data frame, and using the bubble command, we were able to get a spatial plot that tells us about the spatial distribution.

And, we can kind of examine this and see that along this line we see larger circles and as we get away from this line the circles start to get smaller. So, it seems that there is some pattern

and in the next few sessions we will start to explore what this pattern means and how we can characterize this pattern using spatial statistics.

Just a small exercise that I want to leave you with, I want to ask you what type of spatial data is contained in the meuse dot sp spatial data frame. So, try to think about it for yourself for a few minutes and come back and I will give you the answer. The answer is already on the slide. It is a geostatistical data and the reason for that is to remember the geostatistical data is defined on a continuous domain and zinc concentration has a value at every point in the riverbed.

So, the riverbed is a continuous domain, and the type of spatial data depends on the nature of the spatial domain over which the data is defined. So, do not get confused because we seem to have individual points that are the sample from over the continuous domain, but the type of spatial data is always dependent on the nature of the domain over which the data is defined and not the data type in R that we use to store it.

So, you should be able to separate the data type. So, even though we are storing it as a set of points and we have a set of points. In theory, we could have got infinitely many points over the entire riverbed so, because that is a continuous spatial domain. This is an example of geostatistical data.

So, just a summary, we converted ordinary data frames to a spatial data frame and we plotted a spatial distribution, we compared it with the frequency distribution and learned how to use the coordinates command, the bubble command, and the formula operator and became familiar with a spatial points data frame to work with spatial data in R. To explore further I will see you in the next session.

Thank you so much for your attention.