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

Lecture - 47
Session 1

Welcome to another session of the Spatial Statistics and Spatial Econometrics course that is being taught by Dr. Gaurav Arora. My name is Saif and I will work with you today on some practical aspects of this course. Specifically, I will start introducing you to the programming language and the programming environment that we are going to be using in this course to do our practical work with data.

Basically, if we want to excel at spatial statistics or if we want to excel at really any subject, we need two things; first, we need a firm and clear and precise understanding of concepts, principles, and theory. And this is obtained by listening carefully to lectures and questioning the material that you receive and then reflecting on your own.

And another really important thing for understanding is what my teachers call putting pen to paper. So, doing written work on your own, solving the proofs, and the exercises, and doing different versions of them is what refines your understanding of the theoretical material, which you have been exploring with Dr. Gaurav. But there is a second aspect to excellence which is skill-based and this is specifically true in courses like this where there is a computational element.

And by skills what I mean is that you are able to apply your understanding to a variety of real-world problems and real-world data, the amount of data that is available out there today in the public domain regarding all kinds of phenomena and processes is increasing day–by–day and this is a great opportunity to develop the right skills.

And specifically for spatial statistics, the way to develop the skill is by working with this data and actually doing something trying to solve some real problem or answering some specific questions using code and software and trying and failing. So, skill-based development is a little different from developing understanding in the sense that you actually have to do stuff

try and fail, and fail as early as possible because anytime you do anything new, you are going to have failure.

And in coding and programming, it is better to just get the failure out of the way right away and it is ok to fail, it is encouraged to fail. And while working things out on paper is important for skills as well, what is really important is that you actually put your fingers to the keyboard and try stuff out. The best way is to try stuff out, skill development is not so much of course, it is about thinking, but the emphasis is on trying and doing.

And we will work, if you can achieve excellence in both of these things then it aids you in two ways – a, it helps with growth, it opens up new career opportunities for jobs employment, and new opportunities for higher studies and research. And specifically for this course where we are exploring spatial statistics and econometrics, it can also help with personal fulfillment and understanding of the world.

I am sure a lot of you are interested in socio-economic problems of wealth inequality and environmental problems. For example, we will be looking at groundwater depletion in this course, there is air pollution whole host of environmental problems, there are financial problems with depressions, and recessions that affect real people and real lives every single day.

And of course, all of you are familiar with the problems that we have with the pandemic and what underlies all of these problems is spatial relationships and spatial patterns. And if you can develop the understanding and the skill needed to address these problems, you have the opportunity to make a really good social impact and that is personally very fulfilling. We hope that we can help you and explore with you the kind of things you need to do to develop these skills.

So, I will be working with you on practical sessions on tutorials where we will actually be doing stuff, trying, failing, and trying again. So, concretely what will we learn in these sessions what are you going to learn? Broadly you will learn two things while you will learn how to program using a programming language that is called R.

It is just a one-letter name, it is called R, it is a very popular programming language and I will tell you about why it is popular and this programming language provides a whole host of tools for spatial statistics. And related to this is a software called RStudio. So, we will be

using R and RStudio for our programming tasks. And the second component of the practical sessions is the ability to download, explore and find data, the data that you need.

So, you will see some examples, I mean this is not the emphasis of the practical sessions, the practical sessions are really aimed at helping you use R to do spatial stats, but of course, you need data to do that. So, you will see examples of real-world data and I will also show you how to download this data from public government portals and how to bring it into R and then work with it to answer your research questions.

So, I just want to start with a very brief discussion of what R actually is, some of you may be very familiar because it is a popular programming language, but for those of you that are not, we will start from the beginning. So, what is R? R is a free software environment for statistical computing and graphics that is sort of the official definition from their own web page.

And you will notice that it's free. So, it does not cost you anything to use it and it is a software environment. So, it is actually little more than just a programming language, it is a whole environment that includes a programming language, but it also includes other tools and features that you can use to accomplish your statistical computing and to do visualizations and plots.

R has grown actually beyond statistical computing and graphics into many other areas, but really those are the key thrusts of R. So, R is available as free software and it is licensed under the GNU-General Public License which means that it does not cost anything to use and you can also download the source code from the public website.

So, it's free and open source, and that makes it very attractive because if you have the inclination and the technical skill to do so, you can download the base source code of R and make changes on your own. So, it affords a really large amount of transparency.

And it provides functionality for very easy data manipulation. So, you can manipulate a variety of different kinds of data, tabular data, raster data, satellite data, and all kinds of data it has a functionality for you to manipulate it and do calculations and then visualize it into plots and figures.

But one really sort of very important feature of R is that it is not static, it is not just something that somebody wrote and that is all it is, it can be extended the functionality of R, can be extended using something called packages. So, if I care about spatial statistics and spatial econometrics, the core R distribution may not have the functions that I need because spatial statistics is a kind of specialized field.

So, what I can do is I can write an extension package that people can then load into R and then use the functions that I have provided in my package and we will use such packages. So, the fact that R can be extended is one of the main reasons why it is so popular because, over the years people have written a whole range of packages to extend its functionality beyond just data science. And as is expected, there is heavy adoption in the industry and academics.

So, I have just given you a kind of figure on the right here and this figure is showing you the popular data science software and how they are measuring the popularity is that they went to the website indeed dot com and they looked at all of the data science jobs that were posted on indeed dot com, in June on someday 19th June 2017.

So, all of the jobs, that were posted under data science in one day, and then they looked at the skill sets that people were asking for those in the job descriptions. And in those skill sets, they found that R was number 5, it was the 5th most desired skill set on indeed dot com.

And really if you look at the top 5 Python, SQL, Java, Amazon ML, R and C plus plus etcetera, for statistics you only have Amazon ML and R, those are the two. Python and Java are more general-purpose. Python, does it's still computing as well, of course, it is the most popular language.

But SQL and R are kind of different. So, R is amongst the top most desired skills in data science. So, this is just to show you that if you can develop the skill, your employability is expected to rise and increase in value. And this figure is also interesting because it shows you that this is the type of figure that you can make in R, you could have made this figure in R easily.

So you could have gotten this data from indeed dot com analyzed it, done some calculations, calculated the number of jobs asking for a particular skill set, and then graphed it out. So, this figure helps us to understand the demand for skills in the data science industry and these are

the kinds of plots and figures that we will make using R of course, for our spatial statistics tasks.

So, before we start using R, we have to set up our computer to be able to use it and I will show you how to do that in this first session, our goal is basically to just set up our computer to start programming with R, and for this, you need to do three things.

So, I will show you the installation procedure for Windows, but as I have noted here in point number 4, R is also supported on UNIX platforms and MacOS. So, it's across platform software for practicality and constraints. I am only showing you the installation for Windows, if you have another system, if you run a different operating system, you can find the installation notes on their website and I have given the link here in point number 1.

So, we will go through this procedure, we will install the core R distribution which is the main R engine and then we will install RStudio. So, RStudio is something called an integrated development environment for R and I will tell you why it is called an integrated development environment. For brevity what I will say right now is that we are installing this because it makes your job much easier, it makes it easier to use R.

You do not strictly need this, but it makes your job a lot easier and I will be using it for the rest of this course. So, I do encourage you to install this on your system as well. And then once we have done that, today we will actually use our first R command, we will use a command called install dot packages to install a spatial statistics package.

And any R commands or keywords anything that is recognized by R as a function or a keyword, I will use a different font for. So, I will use this blue font. So, whenever you see a slide and you see this then you should know that that is an R-specific keyword. So, let us go through this procedure. So, the first step is, we will install the core R distribution which is available on this page here.

And this is the main, the full name of R is the R project for statistical computing and the website is R hyphen project dot org.

And this is the main home page, feel free to explore it. So, you can download R using this link. So, I already have it downloaded.

And then this is the website for RStudio, rstudio dot com, and then RStudio IDE which is Integrated Development Environment.

You can get many different versions. So, there is the free version and then there is a bunch of paid versions, for this course, you just need the free version. So, go ahead and download this version.

And once you have done this. So, you can pause right now and then go and download these two pieces of software and when you come back you can continue for now, what I will do is I will show you how to install this.

So, I already have it downloaded. So, go in this order, first install R. So, always note the version. So, this is R version 4.2.0 for Windows. So, just double-click this and you can select different languages here, I am just going to leave it at English.

And then this is the GNU General Public License.

So, we can go pass that and then it installs it in program files I am going to leave it at that if you prefer a different directory then you can browse to that directory here.

I would just leave those it takes about 166 megabytes of space which is alright by me.

No, I am just going to accept default startup options.

And then I would like a start menu folder.

And it is going to go ahead and complete the installation here.

So, if that worked out, we can see if we got a R.

So, we can go ahead and install RStudio.

Also in program files, alright go for it.

So, RStudio is a little bigger than R, it is more than twice the size. So, it takes a while. So, now I have installed that, we can try and start RStudio.

So, once you have installed both R and RStudio.

Then what you will do is, you will start Rstudio, you are never going to go and start R by itself. So, if you want to use R, you will open RStudio. And this is what RStudio looks like and now what I want to do is just briefly talk about each part of this window.

So, like I said RStudio is an integrated development environment, and what that means, is that it integrates all of the software components that you need to do your R development. So, this is the console window. So, you can type an R command in here. So, you can say something like you can print something and it prints the output right away.

And a console window is basically if you are familiar with MS-DOS or Linux, it allows you to type commands and see the results right away. So, this is the console window here.

And then on the right, you have something called an environment window in which you will see all of the data and the variables that you have loaded in your project. On the bottom right you have a file explorer which is basically like windows explorer within the R environment where you can browse through all your files and you can see the files that you want to load.

And open those, so, if you plot something, if you write some code to show a figure or some visual then R will allow you to preview that visual in this window.

And the packages tab will tell you all of the packages that you have currently installed in R. So, remember that packages are pieces of software that are written to extend the R functionality, remember we spoke about this.

And then you can ask for help, R has a full fully featured manual and help window and then there are a whole host of other features.

For example, you can open a new R script right here and then you can start typing your code hello R in here and then you can save this as an R file.

So, when you save an R file always save it as dot R, always use this dot R extension. So, you can save this and then you can source it by pressing control s or you can use this menu here, the code menu, you can say source right.

And then you can see when you source, it runs this command that you have written inside the file. So, this allows you to write a series of commands or a function and save it as a file and then it allows you to do that within RStudio. So, this is your editor window where you can

edit your R code. So, we will revisit all of these features, but this is just to show you that if you did not have this, you would have to have separate software for each of these parts.

So, you would need a Notepad editor to edit your code, you would need to run R console to write console commands, you would need to have windows explorer to see your files, and you would need a separate plot viewer, but RStudio allows you to just put everything under one window and do all your tasks together. So, that is why, I highly recommend that you install this. So, for the rest of this course, I am going to assume that all of you have installed R as well as RStudio.

So, if we press control l, we can clear the console, now we have R and Rstudio, and we want to do spatial statistics with R. So, by default, there is no spatial statistics functionality, that is part of R, or if it is there, it is not enough we need something more and there is a lot of packages that are available for spatial statistics. I will just show you today how to install one of them and the one that we want is called gstat.

And so, we were going to run an R command. So, at the console, we will type the command install dot packages. So, it prompts you when you start typing, it gives you helpful prompts. So, you can just select install packages and within double quotes, you have to write the name of the package and the name of the package that we want is gstat.

So, once you have written out your command, you can press enter and it will try and download the package from the internet and then install it into your R environment. So, that happened successfully. So, now we have the package gstat, we can try and look for it in that package.

So, now, we have the package gstat, we can click on this.

And then this is the help files specifically for the gstat package. So, I don't worry these are not going to mean much to you right now, but just to show you that if you click on the package name, it takes you to the help page for that package and shows you the help files and you can click on the function names and look at the help.

So, we are running gstat package version 2.09. So, it is important to be aware of versions because packages sometimes change from one version to the next. So, if something breaks in your code then you know that maybe it was a version change that did that. So, once you

install the gstat package, you have to load it and you can do that by calling the library command, and then you can say library gstat. So, now, it is loaded into your r environment you can start using it. So, I am going to stop here, for today.

And just go back to our summary, and so just to remind you of what we have done today. We have spoken briefly about the R programming language and software environment and what it is capable of and why it is so popular. We have set it up on our computer in a Windows environment by installing R, RStudio, and the gstat package and we have learned very briefly about the different parts of the RStudio development environment.

And after this, I am assuming that you will be comfortable enough to start writing some code which we will do next time. So, see you next time.