Spatial Statistics and Spatial Econometrics

Prof. Gaurav Arora
Department of Social Sciences and Humanities
Indraprastha Institute of Information Technology, Delhi

Lecture - 11A
Stationarity in Spatial Statistics

All right. Welcome back to the 11th lecture on Spatial Statistics and Spatial Econometrics. So, in today's lecture, we are going to sort of start a new topic called Spatial Stationarity. Spatial stationarity is probably the most crucial you know the concept that we will cover in this lecture.

And, you know spatial stationarity is something that we will as analysts, as scientists we will have to argue out or claim scientifically before we can you know go on to even define a statistic like mean or variance or whatever right? So, the idea is that unless we can establish spatial stationarity the mean, the variance and so on and so forth; all of the summary statistics cannot be defined.

Its, so it is the heart of spatial statistics and perhaps the most crucial topic that we will cover in this course ok. So, stationery in stationarity in the context of spatial data now, the first thing that we should probably you know while I introduce what is spatial stationery I will also sort of go over why we need spatial stationarity right? Now, remember that we have seen the figures that you are seeing on the left side of this your screen here, where we have data collected at three locations u_1, u_2, and u_3.

We have discussed earlier that at each location the data are by themselves a realization of a random variable. That is to say that at each location the random variable X, that is the realization X is itself a random variable. So, it can take multiple values with probabilities attached to them. There is also data at location u_2, which has its own you know probability distribution attached. As you know we have a random variable again at location 2 and similarly at location 3 right?

When we look at a realization of data at these three locations, let us say we get a sample of data at these 3 locations. It could be groundwater levels; it could be the coal sample, the soil sample and so on and so forth right. These 3 values, this pair of X of u_1, X of u_2, and X of u_3;

by themselves are a pair of a single realization of a random function that exhibits jointly distributed you know random variables in space right?

So, we had said that random functions are basically jointly distributed you know functions of random variables right? And, any observation that we see is basically, you know, one realization of this random function. On the right-hand side of your screen what you see is a probability density representation which is a bell curve, you know representation on the left-hand side of the screen that is X_1, X_2, and X_3. We have a sort of box plot representation on the right, right?

So, what we see here is that at location u1 the data, the majority of data perhaps 95 percent of data lies within the upper and the lower bars of the box right. So, the distribution of X of u1 would look something like this ok. At location 2, the distribution will look something like this. So, it has a lower variance, it is a tighter distribution, right? Similarly, location X3, well it turns out it's kind of a skewed distribution. There are more outliers on the right-hand side of the distribution. So, I have a situation that looks like this right? And, you know again a much tighter distribution for location 4 right?

Now, at each location, there is a possibility of observing different data points along this distribution with different probabilities. When we get a sample, we get one realization of that distribution which is marked in dark red right? So, this sequence of 4 values that are sampled in space u_1, u_2, u_3, u_4 at these locations are just one realization of each, you know, of the 4 distributions from where the data are drawn on at these locations right?

Now, when we conduct statistics, what we are trying to do is we are trying to use these data right? So, we have sampled let us say groundwater values and we have gotten X of u_1, X of u_2, X of u_3, and X of u_4; the dark red you know dots are where we have realized these values. So, we have these 4 values, but we have to use these 4 values to then infer upon the 4 different distributions of you know in space. If I had 1000 data points, there will be 1000 unique distribution types from where the data are sampled at each location.

All I have is one realization in space at a given location and I want to infer about the entire distribution at this location from this one realization right? Ideally, what I should be able to do is that you know, I should be able to hold everything constant. And, perhaps take 100 or 500 observations from every location right? So, here in dark red you know you see these you know 4 sample points taken at two different, two alternative instances than the dark red realizations right.

So, if for these locations we can sample again and again and again and again; let us say 500 times. Then, I will likely get data that will be representative of each location of the density function from where it was drawn right. But, that is typically not possible, right? It is not possible sometimes due to physiological reasons. Well for example, if you are sampling coal. Once you have dug the hole and you have sampled it out, there is nothing more to sample. So, you really have one data point right?

So, if you are working with geological data where you are basically digging out the earth for samplings right? You basically have to rely on just one sample point at a location to infer the entire distribution at that location right? So, at each location, we have that challenge right. In the case of let us say groundwater values, it can be sampled again and again correctly right?

But, it is not sampled at exactly the same point, for example, you know the central ground water board will probably go out 3 or 4 times in the entire year, you know to sampl groundwaterer from a given well. So, if this is well and you were to sample groundwater values from this well, you know before monsoon and then again after monsoon; the conditions are very very different.

The data that are collected before the the monsoon in a given year are not going to be you know distributed at the with the same distribution are not going to have the same distribution after the monsoon right? So, the distribution that you are trying to infer upon will itself change you know at the second instance. If that happens, you know, effectively we are stuck with the problem that we were facing with coal sampling right?

So, the idea is what do we do? We really have one realization at a location where we want to infer an entire distribution that this realization is drawn from. What we do is assume that you know first of all these data are jointly distributed. So, they are all jointly, there is the space the spatial dimension is sort of binding all, you know, marginal distributions at every location and they are all seen as a joint distribution.

Second, these data are assumed or seen to be stationary that is to say that they are all representing the same distribution right. So, if we can be sure that we are sampling data from the same distribution at each location right? then what we can do is we can pool data across space. And, use that as a proxy for as if those realizations were coming from one location, any one location right? This is the idea of stationarity.

The idea of stationarity is that you know the random function that we are working with, at each location it is exhibiting the same features as every other location right? So, there is stationarity in the sense that I can use the realization of locations 1, 2, 3, 4, 5, 6, 7, and 8 till 100. And, infer upon the density function at location 1 and at location 2 and at location 3 as well as keep going till at location 100 right.

So, stationarity is really about the fact that the data exhibit similar features in that region. Now, quite clearly you know when I am talking about spatial data, we can only ensure you know stationery usually in a given locality. Now, what could be the size of the locality that we will see, you know that we will have to figure out formally as well as you know we will start with some data and try and see what this really means.

But, you know at the end of the day we will have to ensure that we are working with the same distribution at each location. So, then we can pool observations in space because we do not have them through time with everything else held constant ok, alright.

So, we have already talk we have already talked about why do we care about spatial stationarity. Well, we want to build a statistic for the random variable at a given location or at every location that describes certain properties. So, the point is I have one realization at location 1, but I want to infer upon the mean, average realization at location 1. How do I do that right? I will pool the data over space and calculate the mean and allocate it to locations 1 and location 2 and location 3 and location 4.

But, then I assumed that the data are stationary right? To be able to allocate that mean value right to each location, I must be able to claim that the data are stationary in that domain ok. So, the idea is that you know let us say I have these air pollution monitoring stations in New Delhi right? And, at a given time point right on a given day in summer or in winter, let us say on the New Year's, on the 1st of January of 2022, I have one sample of you know PM 2.5 at different locations across Delhi right?

Each location by itself is a random variable right if I am saying that average on average the PM 2.5 concentration in Delhi was about 300 units. Then, what I have done? is I have taken an average across all the stations that I found in Delhi. You see these statistics reported all the time over media and other places right when we are able to commit an average for the entire domain which is let us say the NCT of Delhi.

What we are assuming implicitly is that the data are stationary over the entire Delhi region, right? That they are the distribution that is being represented at each location is the same across that region. Now, whether is it true or not is something for us analysts to figure out? Can we know blindly or you know casually float out such average statistics? Perhaps not, that is not a scientific you know report right?

So, but once we are sort of, you know, trained in spatial statistics or spatial stationarity, we should be able to make such distinctions right? that is the objective of this lecture ok. So, we are really relying on multiple observations over time right? Instead of relying on multiple observations over time, we must pool together samples over space such that the underlying conditions are the same across the data points. And when I say the same, all I am saying is that the data are stationary.

I have also here said the decision to pool data over certain points is the decision of stationarity. You have to be very careful that it is a decision, it is a scientific decision to make to say that it is stationary right? There is not going to be a way to test this statistically right. We are not going to be able to test it using data, we are not going to be able to put it to some kind of empirical you know validation. We really have to make conceptual you know validation of stationary or non-stationary domains.

There is a very nice resource from Professor Michael Pyrcz at the University of Texas. He is a geoscientist and I think he has very good, you know, collection of literature, where he discusses stationarity. I have just sort of taken a poster that was that he shared and I want to sort of go over a couple of things about stationarity and hoping with the hope, that it will make things a little bit more clear.

So, what you are trying to do is you are substituting time for space, something that we have, you know, I have tried to motivate earlier. The idea is that you know, he is giving an example here of coring a sample from space. So, you have a domain and you go on to these locations u_1, u_2, u_3, u_4, u_5, u_6, u_7. So, you have 7 locations and once you have extracted earth out of these locations, you really have a hole.

So, you have no other way to sample that location again right? You will have to probably wait a million years when the depositions will happen and the earth will be filled up naturally. And, then it will probably, probably, or probably not exhibit similar properties right? So, what do we do? What we do is that because we cannot sample at each u_i, again and again, that is over time, we simply pool the observations over space by assuming that the data are stationary, that is they all exhibit the same properties right?

They all exhibit you know properties that are linked to the same probability distribution function PDF or CDF at any given location. And, he gives the geological definition of stationarity. So, he says the rock over the stationarity domain is sourced, deposited, preserved and post-depositionally altered, in a similar manner the domain is mappable. So, you should be able to practically map the domain right? I mean it is not an abstract entity only and may be used for local prediction or as information for analogous locations within the subsurface.

Therefore, it is useful to pool information over this expert-mapped volume of the subsurface right. So, he is talking about a rock deposition. So, with we can think of coal you know samples right? Now, coal which was deposited or you know that came about at any location, at any region in the space perhaps went through the same you know processes to be formed as a block of coal in that region.

And, it is not a single-year process, we know it is a very very long-term process right? So, if we can claim that the coal formation was happened with similar processes you know over time, then perhaps we can claim the area to be stationary. And, pool information over space because we cannot really sample information multiple times you know that at any given location.

A very interesting term that Professor Pyrcz is using is he saying expert mapped volume, that basic that basically you know points to the fact that we need expert domain knowledge to be able to claim stationarity. If I am working with coal data, I need the domain knowledge of mining of rock you know of formation, coal deposition and so on and so forth right?

If I am working with agricultural yield data, I probably should be taking inputs from economists, crop physiologists and so on and so forth, before I can make the claim of stationarity. If I am working with groundwater data, I should perhaps take inputs from hydrologists, you know regional planners and so on and so forth before I can make claim for stationarity, right?

And, he is giving some statistical definitions that you will come to later. But, there are two more comments that he makes first stationary is a decision, it is not a statistical hypothesis that you can test right. So, it is a decision, it is a hard decision oftentimes, but we must take it before we conduct an analysis. The second thing he says is that stationarity assessment depends on scale.

We cannot avoid this decision and you know so, you know we must make this decision right? So, we will look at what it means by the fact that stationary stationarity assessment depends on the scale next ok.

So, to sort of give you an example, I have simply taken a picture of a stone wall on our campus at triple IIIT Delhi, right? So, this picture was taken at the Research and Development building at triple IIIT Delhi ok. And you know so, here is a picture. This is the spatial domain, right? So, we see the spatial domain. It has a wall, it has a metal railing with different colors; you know there is gray, and there is also some you know off-white color.

And, there is also a board that says first floor which is in green color and the question that I am trying to ask is, is this you know domain stationary right? So, now I have to decide as an analyst. So, the first thing that came comes to my mind is that of course, this domain cannot be really stationary. Why? Why cannot it be stationary? Well, it is composed of different materials.

First, we have the stones right on the wall. We have this you know let us say a metal board, which is a completely different material, has a different color different you know spatial properties right? And, we have this metal railing which is again a completely different object. So, perhaps what I should do to find the stationary domain. So, I am basically deciding as an analyst you know, I cannot take all of this; this is a very heterogeneous surface. I have to go and look at a finer location to see if I can find a stationary domain.

So, let us do that. So, I am going to now focus on the area within this red box and try and see if that is you know stationary domain ok. I do that and in this red box what you can see is that again there is this stone wall right? And, also there is this patch, which says the first floor which perhaps is not is sort of the aberration in this quest for a stationary domain.

But, what if I were to sort of you know clip this out and then look at the area? So, can I say that all the region within this metal board is stationary and everything else which is the stone wall is stationary or not?

So, I am going to now focus on this problem again ok. So, I am not only going to look at the stone wall. When I look at the stone wall, what I find is that although I have basically concrete and stones distributed over space right, different stones have different properties. For example, some stones have a very very different color tone to them than others, right?

So, there is, this stone that is sort of light brown, and there is this stone that is grey in color right? There are stones which are you know there is also the very bright white somewhere and so on and so forth. I also see some spatial gradients right. So, there is this little cut where you have this elevation change going up from a higher elevation to a lower elevation right?

So, I still see heterogeneity and I cannot really see all of these observations if I dig out let us say a sample from here and a sample from here and a sample from here, from these three locations. I can in fact; say I can pool this information together and tell you what is happening at each location, it is hard to say that right? So, we cannot really claim stationarity here. But, maybe what we can do is that we can probably focus on one stone on the wall and perhaps claim stationarity.

So, what we are going to do next is we are going to focus on one stone which to the naked eye looks quite a homogeneous surface right? As if, I can just get 10 samples from here, I can at least say what is happening on this you know on this domain on this stone, on this one stone. So, I am going to now focus here. So, at this scale; so, I am kind of zooming in, I am changing scales right.

When I started with the first you know largest scale product that had these, you know meta difference is the stone wall versus the metal board versus the metal railing. The hope is I can go on to one each of these spaces surfaces and maybe you know claim stationarity. When I change my scale and zoom on to the wall only, I sort of think that you know maybe the stones are so different that I probably cannot claim stationary; I should probably look at one stone at a time right? So, I am going to now try and do that with one stone.

When I do that, now I am going to try and you know just draw this boundary to the stone that we are looking at. What you see is a highly zoomed-in picture of a stone right? Now, on this picture you know, which looked quite stationary or homogeneous from a different scale right? When it is zoom when we zoom in further, we can sort of you know see that we have this region, which is much lighter you know seems to be different than the region let us say that it adjoins right.

I can also see a lighter region here; I can probably also see some kind of you know change here and so on and so forth right? So, what happens is as I keep zooming in, my understanding of spatial stationarity or homogeneous surface or something where the underlying processes would be the same, keeps on changing right.

So, if I zoom in enough I can figure out here, I am claiming, I can figure out 5 stationary domains right? I can see that this domain is different from this domain here. There is another you know slightly differentiated domain here and I have a different domain here, right? Maybe, I can even sort of you know figure out, you can figure out more you know more sort of differentiated homogeneous surfaces here, right?

Now, what I am saying is as long as you give me 10 observations in domain 1, I can pool them over space and tell you what is happening on average at a given location right? What if you were to give me you know 10 locations from each of these domains, perhaps it would not be a good idea for me to know pool data across domains. Because there seems to be something different going on in generating data along these domains ok.

So, that is the idea of stationarity, and as sort of you know as Michael Pyrcz suggested in his poster earlier that spatial stationarity assessment indeed depended on a scale right? So, we are going to sort of you know we have established that spatial stationary assessment depends on scale.

So, if I am looking at groundwater data and I am looking at it for a district in Uttar Pradesh versus if I am looking at it for the whole of Uttar Pradesh or if I am looking at it for the whole of India, the stationary assessment will differ ok. So, every time I have data let us say it's groundwater data, but I have it for two different regions; I have to then source the expert domain knowledge to be able to make this decision of stationarity.

Thank you for your attention. See you next time.