Spatial Statistics and Spatial Econometrics

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

Lecture - 11B
Stationarity in Spatial Statistics: Example of Groundwater Data

All right let us look at another example, this example is a little bit more realistic than the stone wall that we saw you know previously right? This example is about groundwater level data in a region. Now, groundwater levels data energy when we think about groundwater, the groundwater issue is very similar to the coal you know seam beneath the surface. Well, what is common is that you know both groundwater and coal are blind to the naked eye right?

So, we cannot really we can when we look down you know we know there is groundwater, but we cannot really know what is the structure within which it exists and what is the flow direction, what are the; what are the different you know dimensions of groundwater, you know in the way it exists beneath the ground.

So, here is a schematic you know a picture that has been sourced from the geological survey and it is a very old picture. You know this picture sort of goes back to 1890, 1896, but this picture is very informative. If you look at what is happening here is that we have the ground surface right.

So, the ground surface by itself is quite an irregular right, and beneath the ground surface, of course, you have this rock structure right, I mean rock or sand structure right? the dotted area. And beneath the dotted area, the rock structure we have is what is called the aquifer, right?

The aquifer by itself is extremely irregular, right? Now, the aquifer the way it exists is that it stores water at times as tubs that are large in size or as tubs that are small in size at different elevations, right? So, if I were to start you know drawing water from this source D and I start drawing water here very soon this water will deplete and I will start seeing the rock or the aquifer surface.

And if I were to solely you know to base my estimation of groundwater levels on location D, then perhaps what I am going to say is that this region does not have a lot of groundwater and most of our most of it has been depleted already. But if I were to go to locations E and F, I will have a very different conception of what is happening with groundwater levels in this region.

So, what this figure is showing us is that the aquifers exist as these tubs that contain water in different volumes and shapes right? When it rains, when it will rain let us say when the precipitation happens, you know what is going to happen is maybe at the elevation the water will start to slide down and probably from here it will start entering the aquifer region and sort of replenish the groundwater right?

At times you know aquifers are also such that water exists you know between hard rocks, right? I mean so this tub here is a domain by itself. So, if you were to draw water from here you would have to really dig down deep you know until here to be able to draw water from here, but location this location A is very very different from this location B well probably A1 and B1, I should be careful about notation here right.

Perhaps they are not stationary, because the source of their replenishment for B1 which is going to come from here, when it went through precipitation for A1 it might be coming from somewhere else, right? So, we should be very very careful before we go on to you know assign stationary domains for groundwater levels, right? So, having sort of seen a picture a very very complex you know perhaps a very complex picture of the way or the structure in which groundwater exists, let us move forward and try and assess you know the quest for stationarity for this surface.

So, the exercise that we are doing here is to, if we are asking can we delineate the factors that impact the spatial stationarity of groundwater data. And we are saying specifically for Uttar Pradesh, India ok. So, I am talking about this region in the pink boundary which is the state of UP in central India and I am trying to see if I can you know delineate the factors that impact the spatial stationarity of groundwater data.

We have seen these data previously when we were doing x, we did an extensive exercise of exploratory data analysis with these data right. So, we have seen the data, we have seen some of the characteristics of these data, you know what are the different levels and so on and so forth. we know that the green dots are representing you know not so deep water depths.

Whereas the color goes from yellow to orange to red, we are looking at you know worsening groundwater level situations, right? Now, the question is where are the stationary domains in these data, is the entire state a stationary domain in itself right? Should I be sort of you know dividing the state you know in let us say 4 parts, 4 equal parts, and key call each of these domains stationarity? Should I divide them you know vertically or should I be dividing them horizontally, right?

What should I do? Quite clearly such a decision should not be arbitrary right, because once I define my stationary domain only then can I say the average you know level of groundwater in that domain. Unless I have claimed or argued out that UP is especially stationary I cannot say what is the average level of groundwater. So, the mean statistic is not defined unless I can claim stationarity, ok.

Now, ok so clearly we need something more than arbitrary you know demarcation of stationary areas. And as you have seen earlier perhaps we need expert domain knowledge and one of the expert domain knowledge is the knowledge of aquifer types. So, one of the things about the aquifer in UP is that it is an alluvial aquifer right, what is an alluvial aquifer? Well, what it means is that there is a very high literal movement of water between these tubs.

So, these rocks are porous in nature, that is water is all the time traveling across these rocks right, if they were non-porous if they were non-conductive then you know the kind the water capacity of what can be stored in these rocks will probably just depend on what is falling down from the surface and what is being extracted for irrigation or domestic purposes or industrial use, right.

But if they are porous, then the water levels are connected everywhere right so the UP region in general is the alluvial aquifer and hence quite porous. That is not to say that you know that then we can just blindly kick that, is a question can we now blindly assume that up by itself is a spatially stationary domain as far as the groundwater data are concerned right? Well, what we will see is probably not because there are other factors to worry about ok.

So, here is another sort of data on an aquifer, you know thickness in meters across UP. So, what we are seeing here is you know so the NCR region is so sorry; so this is New Delhi here. So, what we see is that in the region which is nearer to New Delhi, the aquifer is not so deep right, not so thick so that capacity of storing water is not so high.

But as we move you know eastward there are very very deep aquifers to be found, of course, there are some patches of not-so-deep aquifers in the east location as well, but largely you know moving eastward I have you know much more density of deeper aquifers than I have on the area that is proximate to New Delhi. So, there is a lot of media news of groundwater depletion around Delhi in the NCR region, well a lot of it could be organization demand.

But also, could be that there is less water retention capacity or water storage capacity of aquifers near the Delhi region, alright.

So, based on our discussions earlier, now let us try to identify factors that impact the spatial stationarity of groundwater data, right? So now, we discussed I mean we looked at an aquifer structure. So, aqueous aquifer structures can indeed be very irregular in space, right?

I mean what we have seen till now is that you know you could have the aquifer containers could have different types for example, they could be alluvial rate nature, they could also be clayey in nature. So, if it is clay you know it does not allow too much lateral movement of water, right?

The other thing we saw was what is the thickness of the aquifer, that is to say, what is the water retention capacity of an aquifer right? So, these are spatial you know characteristics of what is going on beneath the surface when we try to understand the groundwater dynamics or how groundwater levels evolve you know beneath the surface right?

So, you know one factor that we are sure of, that we will have to have access to expert knowledge is groundwater hydrology right? Specifically what we have seen here is an aquifer, type. So, we saw that you know this could be alluvial in nature, this could be you know clay in nature, they could be hard rock right there is no movement whatsoever and so on and so forth.

Second, we saw you know a map on the thickness of an aquifer which is nothing but the water storage capacity of an aquifer, there are more factors that groundwater hydrology will guide you with regarding aquifer properties. It is also important you know how fast the water moves from the ground surface, you know let us say if there is precipitation you know how much and how fast will it actually enter the aquifer you know the source of water right?

So, hydrology is important for deciding whether or not a region or a sub-region is a stationary domain for groundwater spatial analysis right. What can be the other factors? Well, one of the things that came up in our discussions just you know 5 minutes ago was that there is there could be different you know structures of demand for groundwater or consumption of them of groundwater right?

Near you know for the UP state, near the New Delhi region right? So, the areas like Goutham Buddha Nagar, areas like Meerut you know and so on and so forth what you expect is a lot of groundwater demand is going to come from domestic use or industrial use right? So, that demand is going to be very different from let us say, if I go into the heartlands you know towards Central Uttar Pradesh or Eastern Uttar Pradesh where a majority of the demand is probably going to come from you know agricultural irrigation of groundwater, using groundwater right.

Now, these two, uses are domestic industrial and are characteristically different from agricultural. Agricultural use is mostly for irrigation, it is also seasonal, it depends on crop cycles you know when you plant a crop, who does plant a crop, and who does not right. So, is a lot of human decision-making going on in terms of the consumption of groundwater in different regions.

So, the second very very important you know factors that will drive whether or not a domain is stationary for groundwater discharge is going to be you know the human dimension or the human impacts right.

These can also be termed the anthropogenic factors of groundwater you know dynamics, right? Well you know you can have exactly the same hydrology beneath the ground, but if you put a farmer on the top or a leather factory or a sugar factory on the top, you can expect very different groundwater level dynamics at those two locations. And we should not treat them as the same domain. So, we should not probably assign them the same average or vary in statistics or whatever right?

So, we should probably have different statistics, a summary statistics for urban regions with a given aquifer type and a separate one for agricultural regions with you know maybe the same aquifer type, right? So, now we are looking at pairs of factors that we should be looking at when we are assigning stationary domains within the Uttar Pradesh region. The third factor is obviously, the rainfall, right?

More generally we can say the weather is right? for again at a given time period if it is good rainfall here while you can expect high this you know recharge, if it is bad rainfall here you know it is a drier patch as far as the weather is concerned then you can expect the water levels to be slightly lower because the demand is sort of you know is ongoing whether or not it rains so much. If it does not rain so much you know for agriculture you know groundwater provides the substitute for you know for cropping needs right?

So, you may have even more vigorous discharge in lower you know rainfall years. So, lower rainfall years will have sort of double whammy for you know agricultural regions as far as the groundwater levels are concerned right? So, certainly, you know different seasons so far as the rainfall levels are concerned you know monsoon, non-monsoon, pre-monsoon, post-monsoon.

Perhaps they should be considered different stationary domains you know to sort of delineate you know summary statistics like mean, variance, standard deviation, the 5th percentile, the 25th percentile, the 75th percentile and so on and so forth ok.

So, what we learn through this exercise of looking at groundwater levels data is that spatial stationarity assessment depends on context right. The first thing we learned was that spatial stationarity assessment depends on scale ok, and the second very important learning is that the spatial stationarity assessment depends on context.

If I were to look at you know cold seam data, oil exploration data, groundwater data, population data, education data you know crime rates data, especially all of these different contacts will involve you know domain knowledge to make the decision of stationarity right. And it will matter, what will also matter is what scale are we making that decision right, what is or what is not a stationary domain depends particularly on these factors ok.

So, it is a complex decision, we cannot avoid it we cannot run away from it. It is the first decision that we must make before we move on to you know signing summary statistics or conducting data analysis or so on and so forth ok.

So, now, I want to move on to the second sort of part of this lecture, which is where I will give you formal definitions of stationarity, a mathematical exposition of stationarity.

The first exposition that I want to talk about is called intrinsic stationarity, right? So, I start with the spatial data model something that we have seen before, what we are looking at is a random variable Z at location s right. So, we have at every location s there are different levels of Z, that we can you know expect to realize and each realization potential realization of Z at location s will happen with a given probability, right?

And s is in a given domain and domain is (Refer Time: 19:07) dimensional space. So, we will always sort of you know simplify our life, and let us say we work with a 2-dimensional domain, ok. So, here I have data at different locations right and you know right and the data at each location are given by random variable Z s, and your small z at this location s can be thought of as a realization of this random variable right and we are you know s is simply the index also so every location is unique.

So, I can just call it location 1, location 2, location 3, location 4 all the way to location n right ok. Then intrinsic stationarity is defined in terms of the first differences ok, and intrinsic stationarity by definition is defined on the first differences. What is the first difference? Well Z at location s and Z at location s plus h right? What is h? h is the spatial lag ok, and h by itself is a vector; that means, it encompasses both you know distance and direction right?

If I have you know, if I am moving from s to s plus h I have the distance as well as the direction. I can have a value of z s plus h prime where the distance, as well as the direction, will be different, right? The property on which you know intrinsic you know stationarity is based is the first difference, the first difference meaning Z s minus Z s plus h prime, Z s minus Z s plus h.

Now, intrinsic stationarity holds when expectation Z s plus h minus expectation Z s is equal to 0, what does it mean? It means that expectation Z s plus h is exactly equal to expectation Z s; that means, the mean value at each location is exactly the same. Remember we are pulling data in space to be able to comment upon summary statistics centrality measures at each location.

So, the first assumption for intrinsic stationarity is that the mean value for Z of s will be exactly the same no matter which location I am looking at, right? So, you know whatever the shape of the distribution, the mean wherever the mean lies will be exactly the same for all locations, right? This is obviously, this is the first-moment property for intrinsic stationarity.

The second-moment property is the variance; obviously, the variance of the first difference is Z s plus h minus Z s. Now, on average that is an expectation that is what we have said in expectation the value Z s plus h and Z s will be the same right? Of course, for a given realization they can be different values, but in expectation, if you were to, if you were able to sort of sample at the same location 1000 times keeping everything else held constant, right?

At the same location same world if you are sampling again on average each location will you will do the same value of groundwater level or the coal quality or the coal quantity whatever you are working with, right whatever the context of your problem right. Now, the difference, the variance, and the difference could be different right it is the second moment. Now, the second-moment property says the variance of Z s plus h minus Z s is equal to what we call 2 gamma h.

So, where the expectation of the first difference was independent of either the value Z s or the lag vector h, the variance of this difference does not matter on the location, but only the distance between the two you know locations s 1 and s 2, right? So, in the second moment of the first difference between these values in space, two pairs of any two pairs will solely depend on the lag between them inside this function gamma h, right?

Now, this 2 gamma h is perhaps the most important parameter of spatial statistics ok, it is called the variogram ok. So, the variogram is, of course, coming from variance, it is a device by itself it is an operator by itself, right? So, the variogram basically is the variance of the first difference of locations, of pairs of locations across space if our given domain of spatial data right?

Now, this variogram only depends on h, it does not depend on Z right, it does not also depend on the location it depends on the distance norm between two given vectors s_1 and s_2 right? You can sort of call you know s plus h as s_2 and s_1 and you know you can sort of think of any two locations in this domain. So, this is true for every s_1 and s_2 pair in domain D which is a subset of 2-dimensional real space, ok alright? So, we have learned the variogram, and we have looked at the property of intrinsic stationarity ok.

Let us move forward. So, I am going to just repeat this one more time, first of all, intrinsic stationarity is defined in terms of first differences. Second of all, it depends on the first-moment and the second-moment operators you know expectation and variance ok. The variance of Z s plus h minus Z s yields is equal to 2 gamma h depends only on h, only on the lag vector, and depends only on spatial lag.

If it depended on locations, that is to say, that it also depended on let us say s, that is which location are you starting at, then it would not be intrinsic stationary ok. So, it can only depend on h ok. Now, stationarity or intrinsic stationarity again is a decision, it is not a hypothesis. I am saying it again and again so that you know we do not fall into the trap of actually trying and testing intrinsic stationarity, ok.

So, intrinsic stationarity is not the only type of stationarity, there are other forms of stationarity in data. So, the second one that we are looking at is called second-order stationarity. What does second-order stationarity mean? Well, let us just draw a domain D again and think about data samples from the locations represented by these crosses in domain D right? Let us say we are looking at location s and location s plus h right, I can also call this s_1 and I can also call this s_2, right?

h by itself is a vector it is the difference between s1 and s2. Remember I am what I am writing is a 2 norm and we have seen in the earlier half of this course that this norm is the distance between you know locations s1 and s2 can be measured as Euclidean distance, at as Manhattan distance as great arc distance and so on and so forth, ok.

So, h again captures both distance and direction ok. Now, you know the first condition for second-order stationarity is that expectation Z of s is equal to a constant for all s in domain D, this is very similar to the first-order condition sorry first order first-moment property of intrinsic stationarity.

This is exactly the same by the way that would mean, this would imply that expectation Z s plus h minus expect Z s which will be nothing but expectation Z s plus h minus expectation Z s; why can I take the expectation operator n? Because it is a linear operator, right? So, it can simply enter n.

Now, both s and s plus h lie in domain D, hence expectation values will be constant at both locations and what I get is just a 0. This is exactly the first-moment property of intrinsic stationarity. The second moment property of entrance of second-order stationarities is different from intrinsic stationarity, it depends on the covariance operator between location Z s_1 and s_2. Remember s_1 is s and s_2 is s1 or s plus h, ok; s plus h ok.

Now, you know what is the difference between the variance of the first difference and the covariance. Now, it turns out they are related. Now, if I were to sort of write down I am going to try and write down the second-order property, the second-moment property sorry; the second-moment property of intrinsic stationarity, is the variance of Z s_1 minus Z s_2 is equal to 2 gamma s_1 minus s_2, where you know previously we had just you know equated this two, this value h right?

Now, this would imply the LHS can be written as a variance of Z of s_1 plus the variance of Z of s_2 minus 2 covariances of Z of s_1 minus Z of s_2. This is equal to nothing but 2 gamma h, where h is simply the distance between s_1 and s_2 right? Now, this is interesting. Now, you basically see that this is my variogram, and here is the term covariance Z s_1. Sorry, it is not minus it is, sorry about this fumble here one second ok, Z of s_1 comma Z of s_2 right, ok.

Now, this term here is what is called the covariogram, which you see right here as the second-moment property of second-order stationarity ok. Now, covariogram and variogram are related and the way they are related is that now variance of Z s_1 is nothing but you can write 2 gamma 0. Because you have you know the distance between these two is 0 right, not sorry no 2 gamma 2 twice of sorry about that ok.

It is covariance with distance 0, this 1 variance of s_2 is also covariance with distance 0, it is just covariance of s Z at s_2 with Z at s_2, right? So, we have a C of 0, C of 0 sitting at both terms here. So, now, you can see that the variogram and the covariogram are linearly related. In fact, they are simply you know inverse to each other, ok. What this means is that if the covariance of look values at two locations is high, the variogram value will be smaller.

That is quite intuitive, well if the distance if the difference between two values well expectation of the values is the same, but if the variation in the disc is the difference of values is high, then you would expect those values to be less related. And hence the covariance will be smaller. If the variance of the difference between two values at different two locations is small; that means, there will be much higher spatial dependence and hence you will have the covariance you know to be high.

So, if the variance of you know Z s1 minus Z s2 is small, then spatial dependence is higher and hence the covariance of Z s1 and Z s2 is also higher ok. Now, this is interesting right, I mean now. So, you have these mirror image properties of the second-order properties defined as an intrinsic and second-order property of stationarity, second-order stationarity. Now, these two definitions only differ by these two factors and these factors appear to be linearly related to each other, just you know linearly negative of each other.

But they are specifically different because covariance is a linear association property right. So, covariance only sort of provides a measure of a linear relationship between the value Z at location 1 and location 2 ok. Variance on the other hand is highly non-linear. So, there are stricter assumptions required to sort of you know assert the covariogram, you know the covariance stationarity or the second-order stationarity in space.

And therefore, intrinsic stationarity is considered to be a stronger property than the second-order stationarity property, right that is why the last sentence on this page says second-order stationarity is also known as weak stationarity or wide sense stationarity in a spatial domain ok, alright?

Let us move forward, there is a strict stationarity strict spatial stationarity property as well, or a strong spatial stationarity property as well and this property relies on the joint distribution itself, right?

Remember when we are working with spatial data, we are working with these random functions, right? So, you have data at location s_1, and you have data at location s_2; s_2 is nothing but s_1 plus h right? And what you are saying here is that you know you have you know the distribution of you know marginal distribution at each location, the marginal distribution at each location they are connected spatially.

So, the data all of these data in this domain are jointly distributed. If these joint distributions remain exactly the same, you know even if we were to move the data by a lag right, from some locations, you know, we sample s_1 to sm and we evaluate the empirical CDF.

Then we move on to a different set of locations again m data points, but with a lag let us say we move 5 steps eastward right and we define an empirical CDF if they are exactly the same, we say the data that is s_1 to sm and s_1 plus h_2 s_1 plus sm plus h which is 2 m values define a stationary, strict stationary or strong stationary domain ok.

Now, we seldom use such a strict stationarity property. Usually, we are working with intrinsic stationarity. So, we will see that most of our you know discussions going forward will rely on this variogram device to measure spatial dependence and to be able to you know model it and then take it to regression analysis for further you know resolving further issues.

So, remember you know in the summary of this lecture, the validity of any summary statistic in space that is the mean, variance, variograms, semi-variogram, covariogram, or whatever depends on your decision of stationarity, we cannot avoid this decision, it is the heart of spatial statistics. I started this lecture with this comment that spatial stationarity is the heart of spatial statistics and I am repeating it. Now, with a lot more knowledge at your disposal.

Now, the decision of stationarity is; however, highly subjective and complicated right? You need to argue it out in a rigorous manner, expert domain knowledge and contextual knowledge are unavoidable when dealing with spatial data, right? Just like any other data sets, you know here we have to be cognizant of expert domain knowledge right? Probably more so in spatial data because of the subjectivity involved, alright.

So, going forward we will be now working on measuring spatial dependence or spatial contiguity from the next lecture. And I hope you had fun in this lecture.

Thank you for your attention see you next time.