Spatial Statistics: Best Practices

Lauren Rosenshein and Lauren Scott present an analytical workflow and teach in detail how spatial pattern analysis and regression tools can be used and applied to your work. They talk about the art and science of finding a proper scale of analysis and introduce tools that will help you accomplish this.

http://video.esri.com/watch/903/spatial-statistics-best-practices

---

Video Transcription

00:01 This is a session. It's called Best Practices. It really builds on the other two sessions.

00:04 So, if you're somebody and you didn't go the Spatial Pattern Analysis class or you didn't go to the Regression Analysis class…

00:09 …we don't want you to leave or anything.

00:11 We just want you to know that…we just want to know that we're going to be going over some stuff…

00:18 …and you're going to feel like we're covering it a little bit quickly because we kind of have those assumptions…

00:23 …and apparently your agendas don't say that…don't say anything about that.

00:28 But, hopefully, you'll still get some good information, and at the end of this workshop…

00:32 …we're going to tell you that there's some resources where you can actually go online and download those workshops.

00:38 So if you didn't see them while you're here, you can see them when you get home.

00:40 It's a free, one-hour download.

00:43 So, introductions. My name's Lauren Scott. This is Lauren Rosenshein.

00:48 And we both work on the software development team on the analysis team in Redlands, California…

00:53 …and the tools that we're going to talk about are tools that we developed.

00:57 We're two Laurens and a Marco, and so we know a little bit about these, and we're very biased.

01:02 We think this, of course, the best thing in the software, but…since sliced bread.

01:10 And the way we're going to do this today is kind of silly. So you'll let us know how this works, too…

01:15 ...is we're going to pretend that Lauren here is a new GIS analyst, and she's tasked with solving a real-world problem...

01:23 ...imagine that.

01:24 So we know she has attended, actually many times, the spatial pattern analysis class...

01:28 ...and she's also attended their regression analysis workshops...

01:32 ...and so we know she has some good ideas on how to proceed with this particular problem...

01:36 ...but we're going to help her all along the way, okay, and give her some suggestions.

01:40 So the context for this analysis is that we have this community that is spending a large portion of its resources...

01:47 ...responding to 911 emergency call data.

01:50 And projections are telling them that their population is probably going to double over the next 10 years.

01:58 So they have...they have some questions, right?

02:00 They have questions like, can we be more efficient in the layout of our police and our fire stations that respond to 911 calls.

02:08 And how effective are the locations that we have?

02:13 We know that some areas of the community get lots of calls.

02:15 Others don't get so much.

02:16 What are some of the factors that promote 911 calls or cause...encourage 911 calls?

02:22 And is there anything that we can do to reduce the number of 911 calls we get?

02:27 Given that the population is going to possibly double, what can we anticipate in terms of the number of 911 calls...

02:35 ...we're going to get in the future so that we can start to gear up for that?

02:38 So this is going to be Lauren's task today.

02:40 And the data she's working with actually is real data from an area near Portland, Oregon. The scenario we made up.

02:55 Oh, my bad. This is because we're tired. I've only...this is my eighth presentation...

03:01 ...but Lauren here 11, plus the Plenary.

03:03 So I do not have an excuse.

03:05 So let's go...the context for the analysis, the population is doubled, the community has questions...

03:11 ...How well are our fire and police located? What are the factors that contribute to high 911 calls?

03:16 Can they be reduced? What can we anticipate given that the population is going to increase? Thank you very much.

03:26 Why am I so confused? Oh, because it came out double-sided. Oh, dear. This is going to be a problem.

03:38 Okay. Right. So, Lauren, one of the things that this community is interested in...

03:44 ...is evaluating the existing locations of their police and their fire units.

03:49 And one strategy that might help them would be to create a hot spot map of their 911 calls...

03:55 ...see where the hot spots for 911 calls are and the cold spots and then we could compare that hot spot map...

04:00 ...to the locations of the fire and police that are actually tasked with responding to them.

04:06 So, just...I know we went over this in the workshop, but just to remind you, the way that the hot spot analysis tool works...

04:12 ...is it looks at each feature within the context of neighboring features.

04:15 And it's looking for statistically significant clusters of high values and...hot spots...

04:22 ...and statistically significant clusters of low values or cold spots.

04:26 And it computes a z-score and a p-value for every single feature that tells you...

04:30 ...if that clustering that is found is statistically significant or not.

04:36 So this is going to be a fun analysis, but there's a couple things that you're going to have to think about.

04:41 The first thing that you're going to have to figure out is what your analysis field is actually going to be.

04:46 Because you're working with incident data, we have a single incident for each 911 call...

04:53 ...and because the hot spot analysis needs an actual count or a rate...

04:57 ...you're going to have to figure out how to aggregate or how to come up with that variable.

05:00 And a couple tools that might actually be helpful to you are the integrate tool and the collect events tool.

05:06 And the second thing that's going to be a little tricky is finding the appropriate scale of analysis.

05:11 The hot spot analysis tool is going to ask you for a distance value.

05:14 So there's a couple tools that might help you find that as well.

05:17 One of them is incremental spatial autocorrelation.

05:19 That's a sample script on our website.

05:21 We'll show you how to download that.

05:23 And the other one is the calculate distance band from neighbor count.

05:28 So, let's go ahead.

05:33 Alright. So, I have to create a hot spot map.

05:38 I've got 911 calls. Each point represents one 911 call.

05:44 So, going to open the hot spot analysis tool, and, like Lauren mentions, I need an input field.

05:54 Oh, how the heck am I going to get an input field. I just have incident data.

05:57 None of these points actually have a value associated with them.

06:00 They're just points with x,y locations.

06:03 I know where they happen but they don't have an analysis field, so what am I going to do?

06:07 Well, something I do very often, which is use the tool help.

06:14 Really. I use the help all the time.

06:20 So, one of the things that is mentioned in the tool help is what this input field is all about…

06:26 …which is it has to contain a variety of values…

06:29 …and that the math for this statistic does require that there be some variation in the variable being analyzed.

06:34 You can't just have all ones, for instance.

06:38 And it says that if I want to use this tool to analyze the spatial pattern of incident data…

06:43 …that I should consider aggregating my data.

06:46 So, I'm going to click on that link to another part of the tool help which actually gives me three methods…

06:53 …three techniques I can use for aggregating my data.

06:56 One involves doing a spatial join to polygons…

07:00 …like census blocks or tracts or whatever kind of maybe geographical unit that you have.

07:06 Another option is to create a fishnet and then do a similar spatial join.

07:12 But Lauren mentioned the method that uses integrate and collect events, so that's the one that I'm going to try to do here.

07:18 And so what is integrate and collect events going to do?

07:22 Well, it's going to first snap features within a specified distance of each other together…

07:27 …so it'll snap them together and then it's going to create a new feature class containing a point at each unique location…

07:35 …with an associated count attribute.

07:37 So that's what collect…

07:38 Integrate's going to snap them together and then collect events is going to count up how many points I have in each location.

07:44 So some will have just one and in some we'll have lots of coincident points where they got snapped together…

07:49 …and then I'll have lots.

07:51 So, I'm going to go in there and I'm going to integrate and I'm going to collect events.

07:56 Now, I happen to have used integrate before, and I'm going to tell you something from one analyst to another…

08:05 …and please write this down or ingrain it in your memory somehow, and I am absolving myself of all responsibility here.

08:15 Integrate actually changes the geometry of your input data.

08:23 There is no output feature class.

08:26 It changes your input data. Okay?

08:30 I cannot say this more enthusiastically.

08:35 Please make a copy of your data before you run integrate.

08:41 Make a copy.

08:43 I know I'm saying it like a crazy person, but if anybody here has ever done this, has this happened to you? Yes. Okay.

08:51 You ruin your data.

08:52 Need to send an e-mail to someone who just sent you data. Please send that data again.

08:58 So, please, just remember you need to copy your data first.

09:01 So you'll notice I have my original 911 call data, and I have a copy of my 911 call data.

09:07 And that is what I'm going to use to integrate.

09:09 So, I'm going to search for my integrate tool. Great way to find tools I don't know where they are.

09:16 And, I'm going to open it up.

09:17 You'll notice there's no output feature class.

09:19 I pointed at my copy, and then I have to pick an x,y tolerance.

09:25 So what that is going to do is it's going to say, how close do I want points to be that I'm going to snap together?

09:32 So with my 911 calls, I know that our…the data that I have for where these 911 calls are coming from isn't really that great necessarily…

09:40 …and a point I could have, I was just talking to Brandon here today, and we were saying, I could have 10 crimes at the 7-Eleven.

09:48 They're all kind of different locations, but they all happened at that same 7-Eleven…

09:52 …and I want to snap those features together so that they're all coincident, they happened in one place.

09:57 So I'm going to use a tolerance in this case that matches kind of like the accuracy of my data.

10:02 So I'm going to say within 30 feet of each other, if points are within 30 feet of each other…

10:07 …they happened in pretty much the same location so I'm going to snap them together.

10:11 I'm going to go ahead and run it.

10:13 It won't look much different. I mean, 30 feet's pretty small, all things considered.

10:18 So the dataset looks pretty similar.

10:20 So now I have a lot of coincident points in here.

10:23 So the next thing that I want to do is use that collect events tool that the help pointed me to.

10:31 I'm going to run it on my copy of 911 data because that data is now integrated.

10:39 I'm going to save my output as my integrated points. Oh, sorry. These are going to be my collected events.

10:52 And, what that's going to do is it's going to go through and for every unique location in my dataset…

10:58 …it's going to see how many points are here.

10:59 Some of them, though, there won't be coincident points. There'll be one. Others, there'll be a bunch of points.

11:05 So it goes from 1 up to 25.

11:07 Now I have a count field…

11:08 …and it's being symbolized, the output of the collect events tool symbolizes those points

based on that count field.

11:14 So now I finally have a count field, and I'm ready to run my hot spot analysis.

11:19 So I go back to the hot spot analysis tool, point at my collected events, use that count field that it created.

11:29 So, the next thing I know…I know I'm going to accept the default for the conceptualization of spatial relationships.

11:34 They mentioned in the spatial pattern analysis workshop that the fixed distance band is a good default…

11:39 …that it keeps the scale of my analysis fixed throughout the whole study area, so that's a good option…

11:43 …but I have to pick a good distance band to use.

11:47 So how am I going to pick the right distance band?

11:49 How am I going to decide how big my neighborhood is?

11:53 Well, Lauren mentioned that there is a sample script out there that I can use to help me find an appropriate scale for my analysis.

12:01 So, where do I get that?

12:04 Well, we are going to point you at the end to this spatial statistics resources page. We'll give you a short URL at the end.

12:12 And here on this page, there's lots of stuff, short videos and tutorials, but there's also these models and script tools…

12:19 …and one of the ones here is this supplementary spatial statistics toolbox for ArcGIS 10, and it has exploratory regression in it…

12:27 …and it has incremental spatial autocorrelation.

12:29 And incremental spatial autocorrelation is what Lauren mentioned might help me find an appropriate scale for my analysis.

12:35 So, if I follow this link, brings me to the geoprocessing model and script tool gallery…

12:40 …and basically all this is is a zip file.

12:43 So I downloaded the zip file, and I unzipped it in a folder here called supplementary spatial statistics.

12:54 You can see there was the zip, here's my folder.

12:56 And in there, there's a toolbox, there's the scripts that that toolbox is calling, and there's some documentation.

13:04 So now, when I go back to my…when I go back to ArcMap in Catalog, if I navigate to that

location on my hard drive here…

13:19 …I'm going to find that documentation, the scripts, and that toolbox.

13:23 And so just that easily, I can now start using those tools right from the Resource Center.

13:30 So now I have this incremental spatial autocorrelation tool, and I'm ready to go.

13:34 One of the things that I can do before I start using it, as I always do before I start using a new tool…

13:40 …I read the documentation. I know, it's hard to believe.

13:49 It's really good documentation, and I am completely unbiased. I didn't write it or anything…

13:57 …and it tells me all about how the tool works. It kind of gives me ideas of some of the things that I'm looking for…

14:02 …with the output of the tool.

14:05 I want to check out that tool, open it up, and there's some similarities to the hot spot analysis tool.

14:12 I need an input field. I still want to use my count.

14:16 And, then I have to choose a beginning distance and a distance increment.

14:20 It's like, are you kidding me? I'm using this tool to choose a distance, and now I have to choose a distance to begin…

14:27 …and a distance to increment with.

14:29 Well, how am I supposed to choose those distances?

14:33 Well, there's tools to help me do that, too, actually.

14:38 It's a never-ending loop. No, it's not. There's an end. I promise.

14:42 So one of the…that's my question. Well how do I choose a good beginning distance?

14:45 Well, there happens to be a little section here called how do I select the beginning distance and the distance increment.

14:52 And it says that a good way is actually to use this calculate distance band from neighbor count tool.

14:58 So, I'm going to look for that tool, calculate distance bands.

15:06 What this tool does is, I point at my collect events, and it's going to tell me the max, the minimum, the average…

15:16 …and the maximum distance at which each feature has one neighbor, right?

15:24 So this is the…the minimum distance is the minimum distance that a feature has one neighbor

at.

15:29 The average is the average distance at which most of the features have about one neighbor.

15:34 The maximum is the, I mean, if you don't use this distance, there's going to be at least one feature that has no neighbors, right?

15:41 There's at least one feature that's 3,598.29 feet away from its next neighbor.

15:49 Now that's a really good distance to use for our beginning distance, because when we do this analysis…

15:54 …we want to make sure that all of our features have at least one neighbor.

15:58 So that beginning distance…the beginning distance, a good distance to use is this maximum one neighbor distance.

16:04 It's about 3,600 feet.

16:08 So, the distance increment…the whole point of incrementing is that we want to keep increasing our neighborhood size, right?

16:15 We're going to find out the intensity of clustering at one neighborhood size…

16:19 …and then we're going to increase the neighborhood size and find out the intensity of clustering there, and so on and so forth.

16:24 So if we increase our distance but no features have any new neighbors, we haven't actually increased our neighborhood size.

16:31 So the goal here is that with each increment, we actually increase the number of neighbors that the features have.

16:38 So a good rule of thumb would then be to use this average one neighbor distance…

16:41 …because that means that most of the features are going to be increasing by about one feature each time we increment.

16:47 So we'll use this 350-foot distance for our distance increment.

16:54 So now we have our 3,600-foot…3,600-foot beginning distance, our 350-foot distance increment…

17:04 …and then it says that I have to create this table if I want to display the results graphically…

17:09 …and I definitely want to display the results graphically, so we'll do that…

17:15 …and it's going go through and what it's doing is it's testing for how intense the clustering is in our data at each one of those scales…

17:26 …starting at 3,600 feet and going up and up based on that increment.

17:30 And what we're looking for is a distance at which the clustering is really intense, those spatial processes are the most pronounced.

17:38 And we can tell how intense the clustering is using the z-scores and p-values.

17:45 So the best way to look at it, we're not going to look for a peak in a bunch of numbers. That's pretty painful.

17:49 So, the tool actually outputs this graph.

17:53 And what I'm looking for, what the help documentation tells me I'm looking for is peaks…

18:00 …because peaks reflect distances where those spatial processes that are promoting clustering are the most pronounced.

18:07 So, in mine, I have a couple peaks.

18:10 The peak that most relates to the question that I am interested in, these neighborhood level hot spots of 911 calls…

18:17 …is this 4,600-foot distance, 4,600 feet, the processes that are promoting spatial clustering are the most pronounced.

18:26 So that's the distance I'm going to use, this peak.

18:29 So, back to my hot spot analysis tool, my 4,600-foot distance, and I am ready to run the tool.

18:38 It's going to go through. It's going to look at each one of my 911 calls…

18:40 …in relation to all the other 911 calls that are within 4,600 feet of me, all the other points within, ah…

18:52 …okay. Sorry, guys. Give me one second.

18:58 I did a very sneaky little thing for a demo I had using Globe, ArcGlobe, the other day. I'm sorry.

19:07 I'm very sneaky. I changed what the default output would be. Don't ever do this. Okay?

19:17 I changed the layer, okay.

19:27 It's a very sneaky demo trick.

19:31 Okay. So now, I'm going to do that again.

19:33 I'm going to rerun it, same parameters, and we're going to see what we're expecting.

19:43 Alright. Normal human-sized points.

19:48 So, we got our hot spots, right?

19:50 We got our hot spots in red.

19:52 Those are areas where we have statistically significant high numbers of 911 calls…

19:56 …and we get the areas in blue where we have statistically significant low numbers of 911 calls.

20:01 And for me as an analyst, I'm kind of happy with this result.

20:04 I totally understand it. I know what the z-scores are and the p-values are.

20:08 Lauren taught me all about that in the spatial pattern analysis session.

20:12 But my boss was not in that session, and he isn't…she isn't going to like…she isn't going to like these points on a map.

20:23 She expecting this heat map that everybody's seeing in the newspapers and on the news and all these heat maps…

20:31 …beautiful, continuous surfaces.

20:35 So how am I going to turn this output, which is a statistically valid output of hot spot analysis…

20:41 …into something that a decision maker might be expecting but still feel really confident in what I'm giving them.

20:49 Well, if I must, I will create a continuous surface out of this output.

20:54 So, I'm going to use a tool that I've used before to interpolate surfaces called IDW, inverse distance waiting interpolation…

21:03 …and it's in the Spatial Analyst toolbox.

21:06 Basically, I'm going to point it at my output, and I'm going to use as a field that it's going to use for interpolation…

21:12 …I'm going to use my z-scores because that's what I want it to use for its interpolation.

21:18 So I'm going to point at the z-scores, I'm going to accept all the other defaults…

21:21 …because, really, this is just for visualization purposes.

21:24 I'm still going to use those true points, that's the true results of the hot spot analysis.

21:32 So it's not going to start out very pretty because they don't know we're making a hot spot map.

21:36 So I'm going to go in here. I'm going to change it to blue to red. My brain is just, okay.

21:45 Blue to red here. Oh, I'm going to use stretched because I think it's prettier.

21:49 Blue to red. My reds are going to be my high z-scores; my blues are going to be my low z-scores.

21:55 I'm going to give it a little bit of transparency.

21:58 And now I have the surface that the decision makers are looking for.

22:07 So, I'm going to leave my points on the map…

22:09 …because I think it's really important that the true output of the hot spot analysis is still there.

22:13 Because at the end of the day, all we can really say is that each one of those points has a z-score.

22:18 That's the output. That's the statistical test.

22:22 I can't say that right here has any particular z-score.

22:24 We didn't do any sort of statistics there. We don't know that.

22:27 But I can create the pretty surfaces if it's going to help the decision makers use the output of my analysis to make decisions.

22:35 So I'm going to leave them both there.

22:38 So at this point, I can look at the map, I've made my hot spot analysis…I've done my hot spot analysis.

22:43 I can see that our response stations…we've got this one response station smack dab in the middle of a hot spot.

22:48 Pretty good location.

22:49 Got this one response station close to one of our other hot spots, close enough to where this hot spot kind of ends…

22:55 …that they're serving that area.

22:56 But then I've got this one response station all the way out in the boonies…

23:00 …and if I'm going to tell the decision makers in my organization which response station I think may be questionable…

23:08 …may be a response station that we might think about really how we're allocating those resources…

23:13 …that's going to be the one that I'm going to suggest we take a stronger look at.

23:18 So I have understood the patterns, mapped them out, helped them make a good decision. Am I done?

23:27 Nice work, oops, nice work.

23:30 But before we go back to the slides, let's think a little bit about your next analysis.

23:36 Whenever I look at a hot spot map, and I know it's the same with you, it makes me ask questions like…

23:41 …why are we seeing so many hot spots in this area over there, and what's going on in these other areas?

23:47 When we see lots of hot spot…lots of 911 calls here, not so many over here, what do you guys think?

23:54 What might be…what might be some of the factors that contribute to lots of 911 calls in some

areas and not so many in others?

24:02 [Audience response] [Inaudible] Population

24:03 We don't have any people; we're probably not going to get a lot of calls.

24:07 We have other ideas?

24:08 [Audience response] Drug and alcohol use.

24:09 Drug and alcohol use, so lifestyle issues. Anybody else?

24:15 [Audience response] Age.

24:16 Age might be an issue, right.

24:19 So there's lots of possibilities here, and actually when Lauren and I looked at this map, too…

24:22 …we thought, I wonder if we are just looking at a hot spot map of population.

24:27 I wonder if we created a hot spot map of population if we would see this same kind of picture.

24:33 And so let's go back to the slides here.

24:41 So notice that we're asking why questions here.

24:43 We're saying, why are we seeing so many 911 calls over here and not so many over there…

24:48 …and from…taking the modeling spatial relationships technical workshop this afternoon…

24:53 …Lauren knows that regression analysis is all about answering these why kinds of questions.

24:58 Like, why are there so many calls over here? Why aren't there so many over there? What might be the factors?

25:04 And, in fact, this is one of the questions that our community is interested in.

25:08 And the way that regression works is it works by modeling a dependent variable.

25:12 In this case, we're going to be modeling the 911 call volumes, as a function of other variables…

25:17 …of other explanatory variables that we think cause or promote or encourage or explain 911 call volumes.

25:26 And as you know very well, Lauren, the most difficult part of regression analysis is…

25:32 …finding that complete set of explanatory variables…

25:34 …finding all of the explanatory variables that are important to whatever you're trying to model.

25:39 And unfortunately, until we find that complete set of explanatory variables, we don't have a properly specified model.

25:46 We don't have a model that we can fully trust.

25:53 And you also probably remember from the regression workshop that you can only fully trust your model…

25:58 …if it meets all of the assumptions of OLS.

26:00 So, let's quickly review what those six checks were.

26:04 You're going to want to find explanatory variables that really are truly helping your model.

26:09 And you know you have good ones if they're statistically significant and if they have the expected sign or the expected relationship.

26:18 You also want to check the variance inflation factor, the VIF values…

26:22 …to make sure that you don't have any kind of problem with redundant variables.

26:27 You want each of your variables to be getting at a different aspect of the 911 call volume story.

26:33 If two of your variables are redundant, it's telling the same story, then your model's not going to be stable…

26:38 …and that's going to be a problem.

26:40 Another really important check is to make sure that the model under- and overpredictions are random noise.

26:46 When you get a properly specified model, it might predict a little high over here; it might predict a little low over there.

26:52 But the under- and overpredictions reflect a spatial pattern that looks like random noise.

26:59 When you have any kind of structure in your under- and overpredictions…

27:02 …it almost always means that you're missing a key explanatory variable.

27:07 And you're also going to want to check the Jarque-Bera diagnostic to make sure that you don't have any kind of…

27:12 …to make sure that the under- and overpredictions are normally distributed, and you don't have any kind of bias in your model.

27:19 If your under- and overpredictions aren't normally distributed…

27:22 …it might mean that you're predicting really well for your low call volumes but not so great for your high call volumes…

27:30 …or you're predicting well in some parts of your study area but maybe not so well in others.

27:35 And, of course, you want a model that performs well, so you're going to be looking at your adjusted R-squared value…

27:40 ...and you're also going to be looking at your Akaike information value.

27:43 You want a high adjusted R-squared value, and you want a small AICc value.

27:50 And we have a tool, and one of the sample scripts, that could really help you find a properly specified model.

27:56 It's called exploratory regression, and the way it works is it tries every combination for a set of candidate explanatory variables.

28:07 If you use this tool, however, you do need to be aware that there's a tradeoff...

28:10 ...and you need to understand the wiggle clause.

28:13 So exploratory regression works by taking this long list of explanatory variable and trying every possible combination...

28:19 ...but it's looking for good models.

28:22 It works a little bit like stepwise regression.

28:24 If people here used stepwise regression, stepwise regression also tries every combination of variables...

28:30 ...but stepwise regression only looks for those models that have a high adjusted R-squared.

28:35 Exploratory regression, this tool, actually tries to find models that meet all of the assumptions of our OLS method.

28:43 But, honestly, even if we run this tool and we don't find a properly specified model, we learn so much about our data...

28:50 ...about the relationships between our variables.

28:54 And now if you have questions or concerned...concerns about exploratory regression, we can definitely talk a lot about this.

29:00 But for now, I want you to just keep in mind that we want to select our candidate exploratory variables very carefully.

29:07 They should be supported by theory, by experts, by common sense.

29:12 And you'll eventually want to come up with some strategy for validating your model if you do use the exploratory regression method.

29:23 Okay, so, why don't you see if you can find a properly specified OLS model for our community?

29:34 So, I think I'm going to completely disregard all of my experience on my thesis and the fact that it took me six months...

29:45 ...and that I never found a properly specific model.

29:49 And, no, I did find some. Don't worry. It's not that bad.

29:53 I just was very unlucky, and I learned a lot.

30:01 But I'm going to be very optimistic, and I think that we're just going to be able to explain 911 calls using population.

30:08 One variable, we'll find a properly specified model, and we will be out of here, off to the party.

30:16 So, I want to test my hypothesis that it's just population.

30:19 And, I have my 911 calls. It's a point dataset.

30:24 I've got tons of data in a set of polygons, my census tracts.

30:30 I've got my population variable.

30:32 I've also got all sorts of other socioeconomic and demographic variables in here, the drug and alcohol use.

30:39 I've got age. I've got all sorts of variables.

30:42 So, how am I going to test my hypothesis?

30:45 Well, in order to jump into a regression analysis, basically what I did is, I aggregated all that 911 call data into my census tracts.

30:54 So now I have a count by census tract of the number of 911 calls.

30:58 The reason I did that is that now I can use all that great census Business Analyst data that I have…

31:06 …and it's all in one dataset, which is what we need for running a regression analysis.

31:10 So, let's jump into that data.

31:15 So now I've got all these polygons that I want to use for my regression analysis.

31:20 They're symbolized by the number of 911 calls, and I want to test my hypothesis.

31:24 And to do that, I'm going to use my ordinary least-squares regression tool.

31:30 I'm going to point it at my 911 call data, give it a unique ID, give it a name, and now my dependent variable is my number of 911 calls.

31:47 This is the thing I'm trying to explain, and my explanatory variable in this case…

31:53 …is that variable that we think we are going to just explain it all with, and that's our population variable.

32:01 And now I'm just going to hit OK. It's going to go through, and it's going to tell us, again…

32:04 …how good of a job we've done explaining the number of 911 calls.

32:11 So, first of all, I'm reminded that I have to check my residuals for spatial autocorrelation.

32:17 So we'll do that, but, first, I'm going to check out how well we did with the other diagnostics

that Lauren just talked about.

32:23 So, first of all, I can see our adjusted R-squared using population is .39.

32:28 We're only explaining about 39 percent of the number of 911 calls, the 911 call volume.

32:33 So, really that's not that great. It's not as good as I was expecting.

32:38 And it's definitely not good enough for my community to make decisions…

32:42 …using that information to project how we're going to be allocating resources…

32:47 …to decide how we're going to try to deal with the problems that we're having.

32:50 So, right off the bat, I'm not very happy with my R-squared.

32:56 I also notice that my Jarque-Bera statistic that Lauren mentioned that talks about model bias…

33:02 …that's statistically significant, and when that's statistically significant, like Lauren said, it means we can't trust our model.

33:09 That's one of those six checks, right?

33:12 So, not a very good R-squared. I already failed the Jarque-Bera test.

33:16 It doesn't…it's not like you can pass three out of the five.

33:19 Each…every one of them needs to be…we need to pass all of them.

33:23 So, I'm already kind of out for the count, but I also want to check my residual map.

33:31 And, frankly, I don't even have to run spatial autocorrelation on these residuals.

33:36 There is a humongous chunk of underpredictions over here, big chunk of overpredictions over there.

33:42 We are…these…our over- and underpredictions are definitely clustered.

33:46 We are not doing a very good job of explaining 911 calls.

33:51 We are definitely missing explanatory variables.

33:56 So, how am I going to figure out which explanatory variables are going to help me get a better model?

34:03 We already kind of brainstormed a bunch of ideas about what variables might be important.

34:08 So, I'm already thinking, I've got a list, I've got a bunch of variables in my dataset.

34:13 Well, even harder than figuring out what explanatory variables…

34:15 …well, what combination of those explanatory variables are going to help me?

34:18 I mean, how many of you guys have run OLS like a hundred times trying to find the right

combination?

34:23 Okay, seriously. I know. Just years you could spend.

34:27 So, you don't have to now, because we have exploratory regression.

34:31 One of the tools I noticed when we were looking at that incremental spatial autocorrelation tool is an exploratory regression tool.

34:38 In that same download, in that same zip file for the supplementary spatial statistics toolbox…

34:45 …we have the exploratory regression tool.

34:46 So that's what we're going to use to try to figure out the combination of variables…

34:50 …that's going to help us explain the number of 911 calls.

34:55 So, I'm going to go through. I'm going to use my 911 call data.

34:59 My dependent variable is the number of calls.

35:01 I think this is my favorite thing about any tool in the whole entire ArcGIS system.

35:07 Just going to click all of the variables that I think might be related to the number of 911 calls.

35:15 It's very rewarding.

35:19 So, we've got about 15 variables here. You could do more.

35:23 But it will take a long time. I mean, it's really testing every single combination of variables.

35:27 These…with just these 15 variables, it's going to try, I think, about 5,000 different combinations right now.

35:34 So, that's another reason why it's a good idea to really think critically about the variables that you're including.

35:40 You don't want just throw everything in the kitchen sink into this kind of analysis.

35:43 It's somewhere in between using the three that you know are right and full out data mining every single variable that's ever existed.

35:54 So, the next thing I have to include is a spatial weights matrix.

35:58 And the reason I need to include a spatial weights matrix is because it's used to run a test for spatial autocorrelation…

36:05 …which is what the side panel help tells me.

36:08 And, it's only going to run that test for spatial autocorrelation on the models that have passed all of the other assumptions of OLS.

36:15 That's really for performance reasons.

36:17 Why test for spatial autocorrelation on a tool that's already failed…on a model that's already

failed the Jarque-Bera statistic?

36:24 So, I'm going to point to a spatial weights matrix that I already have created here.

36:31 And you can do that.

36:32 I got a couple questions about this yesterday of like _____ [unintelligible] the analyst.

36:37 You can create a spatial weights matrix, and there's a tool in the spatial statistics toolbox…

36:41 …called Generate Spatial Weights Matrix.

36:43 So this is where you define how your features are related to each other.

36:46 There's a lot of documentation about that.

36:49 [Audience question] Generate?

36:50 Generate Spatial Weights Matrix. It's in the modeling spatial relationships toolset, right with OLS and GWR.

36:57 Okay. Back to analyst.

37:01 So, then I want to save a report file, and the report file is where it's going to give us all those great diagnostics…

37:08 …that Lauren was talking about, all that really useful information.

37:12 So I'm going to blast, workshop, folder, and I'm going to call it Report 1, and it's going to create that report.

37:24 So the rest of the parameters, actually, these defaults in a search criteria, are really good defaults.

37:32 This is where we decide what it means to be a passing model.

37:36 So, the coefficient p-value, the maximum coefficient p-value by making it .05, it's saying…

37:42 …every single variable in a model has to be statistically significant at the .05 level if it's going to be a passing model.

37:51 So, if you mess around with that…if you mess around with these search criteria…

37:55 …we no longer say it's a passing model that based on the assumptions of OLS.

38:00 It's just a passing model by your criteria, but if you leave these defaults alone…

38:06 …then it is a passing model based on the underlying assumptions of OLS.

38:10 So, I am not going to touch those at all, and I'm going to run it.

38:14 So this is going to go through, and it's testing all models with one explanatory variable, models with two explanatory variables…

38:21 …three explanatory variables, four, and five.

38:24 We're going to not look at it in the Messages window because I know it's pretty small…

38:28 …but we created that text file in our folder called Last Workshop, and I'm going to open up that report.

38:38 It's got the same information. I just get to make it a little bit bigger in my text file, so now we can all look at it.

38:43 So, first we can…we get a summary of all the models that were tried with one explanatory variable.

38:50 We know what our highest R-squares were. We get .56, .55, pretty good using just one variable.

38:56 But, there's this very empty looking area where we're supposed to have passing models, and we don't have any.

39:03 So if we had any models that met all of the assumptions of OLS, they'd be listed there.

39:08 But, I'm not that surprised. One explanatory variable, maybe I'm not expecting to have a passing model.

39:13 So then I have two explanatory variables.

39:15 I'm now I'm up to .74, .7, but still no passing models.

39:20 No passing models with three, but our R-squares are still going up.

39:25 No passing models with four, and no passing models with five.

39:29 So, where our R-square's up to .80, .79, we're explaining a lot of the 911 calls, but we are not meeting all the assumptions of OLS.

39:38 We don't have a passing model, which means we really can't trust that R-square.

39:41 We can't trust any of the output of OLS if we don't have a properly specified model.

39:47 So, why don't we have a properly specified model?

39:51 Is this pretty much all the tool does? Tell me ha, ha, you don't have a passing model.

39:57 It used to do that.

40:01 For real. It did used to do that, and we realized that wasn't very useful…

40:04 …because it's going to happen that you're not going to have a passing model.

40:07 And so we want to give you some useful information about why you don't have a passing model.

40:11 So one of the things that this tool has, I'm noticing right now, is this Global Summary.

40:17 And this summary tells me how…what…which…basically it's going to tell me which criteria are the problem.

40:25 And, it's telling me that for my adjusted R-squared criteria of .5…

40:33 …83 percent of the 5,000 models that were tested passed that R-square criteria.

40:39 So R-square is not our problem. Eighty-three percent of them were above .5.

40:44 For our maximum coefficient p-value, we had 8 percent of our models, about 400 models…

40:49 …where every single one of our variables were statistically significant.

40:52 So that's not our problem.

40:54 Eighty percent passed the multi…the redundancy or the max VIF criteria.

40:59 Ten percent even passed our Jarque-Bera test for model bias.

41:04 But zero percent passed our Moran's IP-value test, and that is our spatial autocorrelation test.

41:11 So, spatial autocorrelation is the problem with our model here.

41:15 That's why we're not finding a good model.

41:17 So at least we have something to go on, right?

41:20 We have a clue about why we're not finding a good model in all of the 4,943 models tried every…

41:27 …they all were…well, actually, all the 22 models where spatial autocorrelation was tried where everything else passed…

41:36 …22 of them passed everything, right?

41:39 None of them met the spatial autocorrelation requirements.

41:42 So, that's a good clue.

41:44 But before I go on to try to figure out what we can do about our spatial autocorrelation problem…

41:50 …I noticed in the documentation, there's some exploratory regression documentation, and I, interpreting the results here…

42:00 …I noticed that there's a lot to this report, a lot more than just that Global Summary.

42:05 There is a Global Summary, yes, but there's also a summary of variable significance.

42:11 There's also a Summary of Multicolinearities.

42:13 So I want to take a look at those before I move on to try to figure out how we can deal with the spatial autocorrelation.

42:20 So going back here, the next thing that we see is this Summary of Variable Significance.

42:24 So what this is telling us is that population, for instance, of all the models that it was used in…

42:33 ...67 percent...67 percent of the time, it was a statistically significant variable.

42:41 Jobs, our jobs variable, 91 percent of the time, that was a significant variable in all the models that it was tried.

42:48 Ninety-five percent of the time, low education was important.

42:52 Our variable for percent vacant, 3 percent of the time it was an important variable.

42:58 So, if this had taken hours if I had used 50 variables, let's say, and this took hours and I didn't get a properly specified model...

43:05 ...and I needed to rerun it with some new ideas, probably not going to include percent vacant again.

43:12 Every single variable I include really increases the number of combinations that are tried, right?

43:17 There's like that exponential increase in the number of combinations that are tried.

43:22 So, any single variable I can rid of is a good one.

43:26 So, this is a helpful way of, number one, figuring out which variables not to include next time...

43:32 ...but, also, just, really, you get a lot of information here about which variables are doing a good job...

43:37 ...explaining the number of 911 calls. Really interesting output.

43:41 Another interesting output is this Summary of Multicolinearity, and this where it's telling us which variables are redundant.

43:48 So, for instance, an interesting one that we found was that our alcohol expenditure variable was in 584 models...

43:56 ...that violated the assumption of multicolinearity, right?

44:01 It was redundant with some variables.

44:03 And 99 percent of the time, it was a college grad's variable. We did not make that up.

44:13 Although it is a welcome comic relief in the middle of this summary report.

44:19 So, this is useful, for instance, if we found this multicolinearity issue...

44:25 ...and one of those variables was really important in the Summary of Variable Significance...

44:29 ...and one of them wasn't quite as important.

44:31 We might use that to help us decide which one not to include moving forward.

44:36 We also just learned a lot of information here, too, about which variables are redundant.

44:41 So at this point, you learned a lot about my data.

44:43 There is some more information in there, and I can learn all about it in the documentation.

44:48 But at this point, I want to find a good model.

44:51 I want to explain 911 calls. That's what my community wants to do.

44:54 So how am I going to do that?

44:55 Well, I know that my problem is spatial autocorrelation.

44:58 That means that my residuals are clustered, my overpredictions and my underpredictions.

45:03 Okay. So how am I going to deal with this problem?

45:07 Well, a really good idea is actually to look at the residuals, right?

45:10 I need to see those residuals, and exploratory regression doesn't create a residual map…

45:17 …because it would create 5,000 residual maps if they did that for every output.

45:24 So I need to run OLS to get a residual map, and I'm going to do that on my…the model here that had the lowest AIC value…

45:32 …the highest R-squared, and that's this model here that has population, jobs, low education, median income, and median age.

45:41 So, I'm going to go in here, and I'm going to run OLS…actually I'll just rerun OLS.

45:51 This time I'm going to use population, jobs, low education, median income, and median age.

46:01 So, we'll go ahead. We're not expecting to see any different diagnostics here.

46:08 These diagnostics are going to match up exactly with what we just saw in the exploratory regression report, right…

46:13 …because we're going to have the same R-squared.

46:16 We're going to have the same variable significance, the same Jarque-Bera.

46:20 All that's going to be the same, because that's what exploratory regression is testing.

46:24 But what we get now is a residual map, and what this helps us figure out is, well, what is going on?

46:31 Why do we have this big red cluster of underpredictions in this area?

46:35 Why aren't we doing a good job predicting in that area?

46:37 What variables might we be missing?

46:39 Because we know that if we have spatial autocorrelation, it means we're missing explanatory variables.

46:45 And looking at the residual map, and this is always true, is a great way to help us figure out what variables we might be missing…

46:52 …especially what spatial variables might we be missing.

46:54 What variable might we include that would get at this spatial structure?

46:59 So, a good way to kind of play around with that is to take a look at what's going on underneath there in our actual basemap.

47:08 So, I'm going to swipe the results…oh, the results of my hot spot analysis and look at the basemap underneath here, some imagery.

47:19 And so what we see is in this area where we're doing a really bad job explaining 911 calls where we've got that big cluster.

47:26 It looks like it's the more urban area of our study area, maybe a more industrial area of our study area.

47:33 We don't really have a variable that's getting at that.

47:36 So, what I did was I created a distance from urban center variable.

47:42 I didn't have any spatial variables in my exploratory regression, which I should have, distance from major highways…

47:48 …distance from urban center, that sort of spatial variables, and if I had, I might not have…

47:55 …I might have had passing models right off the bat.

47:57 But I created a variable that has that data, and, now, I want to try, because I've used my residual map…

48:04 …I want to try rerunning exploratory regression using that dataset that has that variable in it.

48:10 So, I'm going to go back in here. I'm going to rerun exploratory regression…

48:13 …but, instead, I'm going to point it at my…the dataset with my distance variable, turn that variable on there…

48:20 …save this as Report 2, and now when I run it…no pass…oh, we're already with three explanatory variables.

48:30 We're starting to see passing models. Wow. With four, we're getting a ton of passing models.

48:34 Looks like that variable took care of our spatial autocorrelation, and now we have a bunch of passing models.

48:42 So again, I'm going to go to that report, because it's easier to see.

48:47 So immediately I can see I've got a ton of passing models, and I can see in my summary here that now we're passing a bunch of our…

48:58 …we're passing a bunch of our spatial autocorrelation tests, and we're passing…we've got a

bunch that are passing all of them.

49:03 So now I have the horrible dilemma of deciding which of the 50 passing models I have, to use.

49:12 So, you know, you have to think about it.

49:15 Maybe which one has the most remediation implications, right?

49:19 Which one has variables that might help you really make decisions, do something about it.

49:24 I mean, a variable about income is much harder to do something about than a low education variable, right?

49:32 We can't just make people richer.

49:35 But we can maybe help educate people. We have a little more power over that.

49:40 So, another great way to do it is using the AIC value.

49:43 The lower the AIC value, the better, right?

49:46 And just three is a better model.

49:48 So, the AIC value is 681 here. I'm explaining almost 84 percent of the 911 call volume story.

49:56 It involves population, jobs, low education, distance to urban center, and businesses…

50:03 …and, because I used exploratory regression, I know it passes all of the checks of OLS, and I have found the best model.

50:12 Explained the problem and now I'm done.

50:17 Well, almost done.

50:20 There is one thing. We're so close.

50:21 We're so close. This is one diagnostic that you didn't really talk that much about…

50:25 …and that's the _____ [unintelligible] test, the Breusch-Pagan test, and in there it's labeled as the BP, Breusch-Pagan…

50:32 …and we see that that is statistically significant.

50:36 That's not bad news. It means that the relationships that we're trying to model do vary across our study area…

50:42 …and that's not bad news, because we do know that we have a properly specified OLS model.

50:46 Sometimes it's good news.

50:48 Sometimes it means that we can actually improve our model by moving to a method like geographically weighted regression…

50:53 …that allows those variables to change over our study area.

51:00 So, let's go ahead and see if, remember what the adjusted R-squared value is for our best model and the AIC value…

51:07 …from this model and then let's see if GWR actually can improve our results, and if it does improve our results…

51:12 …we're definitely going to want to create some maps with the coefficient services because the…if the coefficients…

51:21 …if our explanatory variables have remediation implications, we can use those coefficient services actually…

51:26 …to design targeted interventions or remediation.

51:30 Okay. So, now we're going to run GWR, but just like we talked about in our modeling spatial relationships workshop…

51:46 …the hard part is finding a properly specified OLS model.

51:49 And we've kind of dug into the ta-da moment of finding a properly specified model.

51:54 It's not as bad as it seems, especially with exploratory regression, but once we find that properly specified model…

52:00 …I can just use that when I move to GWR.

52:02 So I don't have to find a properly specified GWR model.

52:06 I found it in OLS. I'm ready for GWR.

52:09 So I'm going to use geographically weighted regression.

52:13 I'm going to point it at my data, same dependent variable. It's my calls.

52:19 And then I want to use my population, jobs, low education, distance to urban center, and businesses.

52:31 And I'm going to use an adaptive kernel type to use the number of neighbors instead of a fixed distance.

52:37 I'm going to go ahead and run it. We're going to see if it improves our model.

52:41 Oh, first of all, I see our adjusted R-squared is up from .83 to .86.

52:48 So we've improved the model there.

52:51 Our AIC value, does anyone remember what the AIC value was?

52:56 [Audience response] 681.

52:57 681? Okay. So, we've gone from 681 to 677, more than 3 decrease.

53:03 So, that means that we have significantly improved our model using GWR.

53:09 So that's awesome, because it means that we have a model that we can…that's even better than before…

53:14 …and it's a great model to use if we want to make predictions or if we want to use this to actually remediate.

53:22 But, the best thing about GWR…well, I don't want to say that…but one of the best things about GWR…

53:28 …is that we can look at the coefficient surfaces.

53:30 So we can go in and look at how, for instance, the jobs variable varies across space.

53:40 So, we can see that it's in the darker red areas that the jobs variable has a very strong relationship with the number of 911 calls…

53:51 …or we can look at coefficient number three, which is that low education variable, and we can see that…

53:56 …these are the areas where our low education variable has a very strong relationship with the number of 911 calls.

54:03 Those are the areas where, if we wanted to implement some change in regards to education, that we would, I guess…

54:12 …get the most bang for our buck, right?

54:14 Yeah, and this is pretty important because suppose that this community decides as a way to hopefully reduce the number of 911 calls…

54:21 …but also to just make some improvements overall to the community.

54:25 Suppose they decide to set up a program, a stay-in-school program.

54:30 Well, they could implement that program everywhere, but if resources are limited, and resources are always limited…

54:36 …they might want to at least try rolling out that program in those areas…

54:40 …where the education variable was the strongest predictor of 911 emergency calls.

54:46 And there's one more analysis that we can probably do to help out our community, and we can do that with GWR…

54:52 …and that is that we do know something about what the projected population's going to be.

54:57 And we can hopefully predict…use GWR to predict how that's going to influence the 911 call volumes down the road.

55:06 Okay. So, we can use GWR. This is really easy, actually.

55:14 I'm going to use the exact same run. It's going to use the same five variables.

55:18 It's going to calibrate the model the exact same way.

55:22 And like Lauren talked about in the modeling spatial relationships section, I can point at a dataset that has my predictive variable…

55:31 …in this case, it's the same dataset.

55:33 And now, I can use my future population and then my jobs, my low education, my distance to urban center, and my businesses.

55:44 And ideally, all of these would be projected datasets if I had them…

55:48 …but just for the purposes of getting a rough estimate of just population alone, if just the population increased, what would that…

55:56 …how would that impact our number of 911 calls?

56:00 And then we'll put our output feature class here.

56:07 And, we'll run it and it's going to go through.

56:11 We're going to see the exact same R-squared. We're going to see the exact same AIC value…

56:15 …because it's using the same variables to calibrate it.

56:19 But now we can go in here and symbolize this using…the same way that we're symbolizing our 911 call data…

56:28 …except now, instead of the actual number of 911 calls, we're going to use our predicted number of 911 calls.

56:35 And we'll go in here and we can see, when we compare that to our original number of 911 calls…

56:43 …that we swipe back our predictions, we can see, I'll start with the original.

56:51 So here's our original, and here are our predictions, and actually, let me…yeah, I turned on the one that's not…

57:08 Transparent.

57:09 Transparent which isn't very useful for comparison purposes.

57:12 So, here's our original data. These are the number of 911 calls, and we can see that if our population doubles…

57:19 …we're going to expect a bunch more 911 calls in quite a few, pretty much everywhere's getting a little darker.

57:25 We can see the areas where we're expecting it to get…we're expecting the change to be the most severe.

57:31 So we can make a prediction and help them do some planning for the future.

57:36 Yep. That's great. I think we are done.

57:39 Let me summarize what you did with your analysis.

57:43 Okay, so Lauren used hot spot analysis so that the decision makers could evaluate how well…

57:49 …the fire and police stations were located, and maybe adding new police stations or moving new police stations…

57:55 …would help them be more efficient.

57:57 She then used OLS to try to identify what are those key factors that contribute to 911 call volumes.

58:04 And where those factors had remediation implications, her GWR analysis of the coefficient values suggested…

58:12 …where those projects or programs might be most effective or where they might have their biggest impact.

58:18 And finally, Lauren could use GWR to predict the volumes, what we anticipate the call volumes are going to be in the future.

58:25 And this not only helps the community to anticipate what they…they're going to expect in terms of 911 call volumes…

58:32 …but it also provides a yardstick for measuring how effective any of those programs might be.

58:41 And before you go, we want you to know that if you're interested in spatial statistics…

58:45 …and any of the stuff that we're talking today in this course kind of went through the pattern analysis and the regression analysis…

58:51 …that we have this web…this resources page that we keep up to date, esriurl.com/spatialstats.

58:59 We have short videos here. We have articles. We have online documentation.

59:03 There's actually free Virtual Campus courses that you can download, so if you missed the pattern analysis…

59:09 …the regression analysis course, you can download them here for free and watch them.

59:14 And also, we put our e-mail addresses here. We hope you will consider us resources as well.

59:19 And if you can, we would appreciate feedback.

59:22 This is the first time we're doing this course, and if you have feedback on how we could improve that, we'd be very grateful.

59:28 And I think you guys are awesome for being here at 4:30 on Thursday afternoon of a very long week…

59:34 …but I hope that you guys had an awesome 2011 User Conference.

59:40 I know that we did, and thank you, and if you have any questions, we will take them.

59:51 [Inaudible audience comment]

59:52 I understand. But go ahead, yeah. We've eaten it. We will take some questions. We've got time.

59:59 [Audience question] Just a question on validating _____ [inaudible]. Any suggestions _____ [inaudible]?

1:00:10 Yeah. For validating your model, that's a good point, is when we do find our models using exploratory regression technique…

1:00:19 …we do want to think about validating, because it reduces our ability to make inferences. That's one of those tradeoffs.

1:00:25 And so, one of the ways that you can validate it is to hold 25 percent of your data aside…

1:00:32 …build your model with 70 to 75 percent of your data, and then make sure that it still works on that 25 percent that you held aside.

1:00:39 In some cases like this one, we want to run geographically weighted regression…

1:00:44 …we have very close to what the minimum number of features are.

1:00:47 So in this case, and I worked on a malaria paper where I did…took this approach.

1:00:52 I didn't have enough features to really hold some aside, and so what I did is, I created…resampled a hundred times…

1:00:59 …and I want to see in all hundred of those samples that the variables that came up as being significant are still significant.

1:01:07 And so that's another way that you might want to validate the model.

1:01:19 Um-hmm.

1:01:20 That second file you already had named? How did you create that _____ [inaudible]?

1:01:25 Wait. Say that again.

1:01:26 You had created an extra variable that was different from the urban center. 01:01:29

1:01:30 It was already a layer.

1:01:31 Yeah. So for the distance from urban centers variable, what I did to create that, I used the Near tool.

1:01:36 Okay.

1:01:37 Basically created a point at the urban…the center of that urban area, and I did a Near from that point to every feature in my dataset…

1:01:47 …and it basically creates a new field with that distance, and you'll…every time we start a new analysis…

1:01:55 ...I get sick of that Near tool because I spend about a day just calculating the distance from every feature in my dataset...

1:02:01 ...to about everything I can think of to create those spatial variables.

1:02:04 So that's a step you'll kind of get really used to by the end of it.

1:02:10 [Audience question] So, once you added that distance variable and there were no longer any stationarity issues...

1:02:17 ...what made you decide to use to use GWR, other than for demonstrations purposes here? Why would you...

1:02:24 So the, so when we ran...when we added that variable, we didn't deal with the nonstationarity issue.

1:02:33 It was actually with the spatial autocorrelation issue.

1:02:36 And the nonstationarity we...we knew we had nonstationarity because of the Breusch-Pagan test which is also there...

1:02:43 ...and that's why we went to GWR because that's statistically significant, which doesn't mean you don't have a good...

1:02:48 ...you can't trust your model.

1:02:49 You still can trust your model if you have nonstationarity, because we have...our tests are robust to that nonstationarity.

1:02:56 It just means that you might improve NGWR which is why we moved to GWR.

1:03:03 [Audience question] Will this session also be posted on that page later?

1:03:06 I have plans in my head of recording this session. We're going to do it.

1:03:11 So, we just have to do it.

1:03:13 I think they are recording this. I'm not sure how that works.

1:03:16 But, chances are because we are a little bit crazy, we won't be that in love with how we've done it today, or...

1:03:25 ...we'll have to write out every word we're going to say and then make sure it's perfect, because we want to make sure...

1:03:29 ...that it's perfect for you guys. So, as soon as we do that, yeah, we'll put it up there.

1:03:38 [Audience question] You mentioned the VIF stability have variables that are too similar, is that correct?

1:03:46 So what do you do? Take out...

1:03:48 Exactly. So, when you have multicolinearity in a regression model and you see that...

1:03:55 ...you're never going to have just one variable that has a high VIF value.

1:03:58 In any given model, you're going to have at least two that have that high VIF value because it means that they're redundant.

1:04:05 And so, what we usually do is to get rid of one and try, put it back in, and get rid of the other one and see which one is better.

1:04:12 And usually, it's based on theory.

1:04:15 So if this a better...has better remediation purposes, or theory supports this one but not so much the other one...

1:04:20 ...then that's how you decide which one to remove.

1:04:23 And sometimes it's actually not just two variables but three variables are telling the same story together.

1:04:30 [Audience question] When you _____ [inaudible] to your client, they say, why didn't your _____ [inaudible] just explain _____ [inaudible] or something like that.

1:04:39 Yep.

1:04:44 [Audience question] Hello. Imagine for this example, imagine instead of...in the example you were using...

1:04:49 ...a _____ [inaudible] check on like everything _____ [inaudibl]e. Now imagine those were, in this example...

1:04:56 ...police beats, something like that, and you're giving this information say, to the chief.

1:05:01 The chief wants to use this information to determine whether or not he needs to redraw those beats. Can you do that?

1:05:14 _____ [inaudible]. In my line of work, you know, a lot of other situations are seen and a lot of scenarios where...

1:05:20 ...they'd like to see, especially since you're using 911 calls, in the police department, they like to see if there's a...

1:05:26 ...high preponderance of calls here, can't just publicly _____ [inaudible] to the police station, but you may need to readjust your beats...

1:05:32 Right.

1:05:33 for that control to be more effective in certain areas. Maybe there's too many calls in this one area to cut the beat off.

1:05:38 Right.

1:05:39 And if there's not enough in this area, you can consolidate others. So I was just wondering how you just _____ [inaudible]...

1:05:45 ...tips on maybe how to take that into account in your statistic analysis.

1:05:50 Oh, just off the…what I think when I hear that is that, that would be more related to like the output of the hot spot analysis…

1:05:58 …than to the regression analysis. I would be looking at how that hot spot map matches up with my police beats and say…

1:06:04 …oh, gosh, I have one police beat covering basically the whole hot spot and 10 that are all cut up…

1:06:09 …and maybe start moving some of those other beats into that hot spot so that they can kind of share the load based on that.

1:06:19 Because it sounds more like a question about the spatial pattern of the 911 calls than about explaining what's causing them.

1:06:26 That would actually be a really good idea. And another possibility is, it does sound…

1:06:31 So, so let's try this scenario. So I know that the population's going to double in 10 years.

1:06:36 I can predict what the 911 calls are going to be, and I can see, oh, gosh, this police beat or the crime's going to be.

1:06:43 I can see that the crime in this police beat's going to be five times more than every other one…

1:06:48 …so my sergeant might say, yeah, you know what, I need to redraw these.

1:06:52 And the hot spot analysis would be one way to do that.

1:06:55 Another would probably be some kind of allocation analysis to try to allocate crimes to make them more efficient.

1:07:06 Yeah, like a location allocation type of thing.

1:07:09 [Audience question] Couldn't you take the results of like what you did and you just allocate it to the police beat polygons…

1:07:14 …rather than _____ [inaudible]?

1:07:16 Sure. For a regression analysis, you could absolutely use the police beats for your regression analysis.

1:07:21 But I still think for…and that might be a better way to do it if that's how you're allocating resources…

1:07:28 …then you would want to be analyzing your data at the scale that you're allocating resources.

1:07:33 But, if you're question is, how do I design those police beats, I think that looking at like a hot spot map…

1:07:39 …or analyzing the patterns of the 911 calls would be more useful.

1:07:48 [Audience question] What can we do for you? It's a serious question. A lot of us think that

what you're doing is…

1:07:56 …the best thing in ArcGIS, or close to it, you know.

1:08:00 We did not pay that man.

1:08:04 He doesn't work for us, I promise.

1:08:06 [Inaudible audience comment]

1:08:09 [Audience question] What can we do to see the Esri since all the resources _____ [inaudible]?

1:08:16 So one place is the session evaluations. That's a really good place. We really appreciate that. Thank you very much.

1:08:25 And sometimes it is hard for us to get data, so if you guys…because our job is actually developing the software…

1:08:33 …so, you know, we wish we had way more time to do real research, to do analysis, and we just don't get a lot of time to do that.

1:08:40 So sometimes we would love if you have a data or interesting analysis.

1:08:45 We love talking to you about the kinds of things that you're solving. That's like the best part of our jobs.

1:08:50 We often when we do that don't sleep.

1:08:53 You know, it's like we have way too much on our plates.

1:08:56 So anybody who wants to mention they need more staff…

1:08:59 Yeah, that would be good though, but…

1:09:01 You could spread that rumor.

1:09:02 Yeah, and if you want to see your data up here next year, because we're dying to redo our tech workshop demos for this year…

1:09:10 …but then we had three new tools and they got up on the main stage and so did I, which is half of our demo creation team…

1:09:19 …so that didn't happen this year.

1:09:21 But, you know, we're always looking for new data to analyze for good examples so that, I mean…

1:09:27 …because…you know, we don't analyze your data to publish it in papers.

1:09:31 We do it to help everybody learn how to be successful using the tools, so…

1:09:35 Yeah. And it can help you, too. I mean, we can maybe have some suggestions for how to do this.

1:09:39 Yeah, we'll analyze the…

1:09:41 We would love to redo all our sessions. We just…every year that's our goal.

1:09:45 So, we're going to, you know, change this and puppy up, and then the reality sets in and we have like no time, so…

1:09:54 [Audience question] So have you used any of these tools that you've done lately _____ [inaudible] thesis _____ [inaudible]…

1:09:58 …to see if there's any improvement?

1:10:00 Actually, the exploratory regression tool was built basically…I mean, we had a precursor to it, but it…

1:10:06 …the tool as it is now was built as I asked questions through my thesis. Oh, what if I had a summary of this?

1:10:14 Oh, what if it was like this? And so that's what it is.

1:10:18 [Audience comment] Better keep taking classes.

1:10:22 Stay in school.

1:10:25 [Audience question] How well do your tools work on data that's in a geographic projection…

1:10:29 …or does it need a _____ [inaudible]?

1:10:30 That's a great question. That's a really, really good question.

1:10:34 You should always be using projected data when you're using the spatial statistics tools…

1:10:39 …because they are using, most of the time, some sort of a distance calculation, and that distance, if it's not projected…

1:10:47 …is just going to be flat out wrong.

1:10:49 So, the better of a projection you can get, the more accurate you can get those distances, the better off you are.

1:10:55 [Audience question] So it's really the distance function that we should caring about so _____ [inaudible]…

1:10:58 …that is the constant distance not a constant area or _____ [inaudible].

1:11:03 You know, that…so, what do we use? We always use something that has Albers' name in it.

1:11:08 Yeah, Albers.

1:11:10 And also equal…

1:11:12 Well, so it's for like the whole United States, we might use Albers' equal-area, but, I mean, yeah, or an equidistant…

1:11:20 [Inaudible audience question]

1:11:21 Some…I mean conic is the other thing.

1:11:24 But, as much as we can…Albers' conic…equidistance.

1:11:28 [Audience question] Equidistance is the tool.

1:11:30 Yes. Okay. But, I mean, especially like a dataset like the ones we were just looking at.

1:11:35 The more you can get into like a UTM, I mean, the closer you can get to the true…

1:11:40 …the best projection for the dataset that you're working on, the better.

1:11:44 [Audience question] I just think there's a Northern Hemisphere set out there.

1:11:46 Oh, then, yeah.

1:11:49 Yeah, you're going to have, I mean, even…I mean no matter what you do, that's going to be a problem.

1:11:53 [Audience comment] Yeah.

1:11:55 Because nothing's going to be perfect, but, yeah, it's definitely important if you use geographic coordinate system…

1:12:01 …for the Northern Hemisphere, you're going to have trouble, and you won't even know it.

1:12:06 Although we will throw a warning saying you should be using projected data.

1:12:09 Yeah. But it will work, just a warning.

1:12:14 [Audience question] _____ [inaudible] article how do you like us to refer to _____ [inaudible], like to refer to a specific article or _____ inaudible.

1:12:25 Our stuff?

1:12:27 Well, for instance, if you're referring to like the tools themselves and you say, the spatial statistics tools in the ArcGIS…

1:12:37 …in ArcGIS, that's what I've done in my papers.

1:12:42 As far as referring to like our help documentation or our ArcUser articles and that stuff, most of the time, you know…

1:12:49 …we're…we've got resources that we will point people to, the kind of the source material…

1:12:59 …for a lot of the stuff that we're talking about, like the Gettis-Ord paper for hot spot analysis, for instance.

1:13:08 Yeah, for all the tools, there's a section that says Learn More About, and I could be missing a few…

1:13:13 …but hopefully at the end of every one of those are articles that talk about the seminal

article for that statistic and…

1:13:20 …other valuable articles that you can cite also in your research.

1:13:23 And also, Marco has at the beginning of a lot of the tools in the code itself, what paper he used when developing the tool…

1:13:31 …what algorithms he used for developing the tool, so that's another place you can go to figure out where he got that math

1:13:41 _____ [inaudible]. It's over. Thanks, guys.

© Esri 2013 http://www.esri.com