

# Understanding Spatial Statistics in ArcGIS 9

## Transcript

Copyright © 2006 ESRI

All rights reserved.

The information contained in this document is the exclusive property of ESRI. This work is protected under United States copyright law and other international copyright treaties and conventions. No part of this work may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, or by any information storage or retrieval system, except as expressly permitted in writing by ESRI. All requests should be sent to Attention: Contracts and Legal Services Manager, ESRI, 380 New York Street, Redlands, CA 92373-8100, USA.

The information contained in this document is subject to change without notice.

@esri.com, 3D Analyst, ADF, AML, ARC/INFO, ArcAtlas, ArcCAD, ArcCatalog, ArcCOGO, ArcData, ArcDoc, ArcEdit, ArcEditor, ArcEurope, ArcExplorer, ArcExpress, ArcFM, ArcGIS, ArcGlobe, ArcGrid, ArcIMS, ArcInfo Librarian, ArcInfo, ArcInfo-Professional GIS, ArcInfo-The World's GIS, ArcLocation, ArcLogistics, ArcMap, ArcNetwork, ArcNews, ArcObjects, ArcOpen, ArcPad, ArcPlot, ArcPress, ArcQuest, ArcReader, ArcScan, ArcScene, ArcSchool, ArcSDE, ArcSdl, ArcStorm, ArcSurvey, ArcTIN, ArcToolbox, ArcTools, ArcUSA, ArcUser, ArcView, ArcVoyager, ArcWatch, ArcWeb, ArcWorld, Atlas GIS, AtlasWare, Avenue, BusinessMAP, Database Integrator, DBI Kit, ESRI, ESRI-Team GIS, ESRI-The GIS Company, ESRI-The GIS People, FormEdit, Geographic Design System, Geography Matters, Geography Network, GIS by ESRI, GIS Day, GIS for Everyone, GISData Server, InsiteMAP, JTX, MapBeans, MapCafé, MapObjects, ModelBuilder, MOLE, NetEngine, PC ARC/INFO, PC ARCPLOT, PC ARCSHELL, PC DATA CONVERSION, PC STARTER KIT, PC TABLES, PC ARCCREDIT, PC NETWORK, PC OVERLAY, PLTS, Rent-a-Tech, RouteMAP, SDE, SML, Spatial Database Engine, StreetEditor, StreetMap, TABLES, the ARC/INFO logo, the ArcCAD logo, the ArcCAD WorkBench logo, the ArcCOGO logo, the ArcData logo, the ArcData Online logo, the ArcEdit logo, the ArcExplorer logo, the ArcExpress logo, the ArcFM logo, the ArcFM Viewer logo, the ArcGIS logo, the ArcGrid logo, the ArcIMS logo, the ArcInfo logo, the ArcLogistics Route logo, the ArcNetwork logo, the ArcPad logo, the ArcPlot logo, the ArcPress for ArcView logo, the ArcPress logo, the ArcScan logo, the ArcScene logo, the ArcSDE CAD Client logo, the ArcSDE logo, the ArcStorm logo, the ArcTIN logo, the ArcTools logo, the ArcView 3D Analyst logo, the ArcView Business Analyst logo, the ArcView Data Publisher logo, the ArcView GIS logo, the ArcView Image Analysis logo, the ArcView Internet Map Server logo, the ArcView logo, the ArcView Network Analyst logo, the ArcView Spatial Analyst logo, the ArcView StreetMap 2000 logo, the ArcView StreetMap logo, the ArcView Tracking Analyst logo, the Atlas GIS logo, the Avenue logo, the BusinessMAP logo, the Data Automation Kit logo, the ESRI ArcAtlas Data logo, the ESRI ArcEurope Data logo, the ESRI ArcScene Data logo, the ESRI ArcUSA Data logo, the ESRI ArcWorld Data logo, the ESRI Digital Chart of the World Data logo, the ESRI globe logo, the ESRI Press logo, the Geography Network logo, the MapCafé logo, the MapObjects Internet Map Server logo, the MapObjects logo, the MOLE logo, the NetEngine logo, the PC ARC/INFO logo, the Production Line Tool Set logo, the RouteMAP IMS logo, the RouteMAP logo, the SDE logo, The World's Leading Desktop GIS, Water Writes, www.esri.com, www.geographynetwork.com, www.gisday.com, and Your Personal Geographic Information System are trademarks, registered trademarks, or service marks of ESRI in the United States, the European Community, or certain other jurisdictions.

Other companies and products mentioned herein are trademarks or registered trademarks of their respective trademark owners.

# Understanding Spatial Statistics in ArcGIS 9

**Presenter: Sandi Schaefer**  
**ESRI Washington, DC**  
**Co-Presenter: Dr. Lauren Scott**  
**ESRI Redlands, CA**



Hi, my name is Sandi Schaefer and I am an instructor with the Educational Services team in Washington DC. Joining me today is Dr. Lauren Scott from the Geoprocessing team in Redlands. We would like to welcome you to the live training seminar on Understanding Spatial Statistics in ArcGIS 9.

## Seminar overview

---

### ➤ Topics

- Basics of spatial statistics
- Measuring spatial distribution
- Spatial pattern analysis

### ➤ Format

- Each topic followed by a software demonstration, review, and Q & A session.

In today's seminar we will discuss three main topics.

First, we will talk about some basic concepts of spatial statistics. Then we will show you ways that you can explore your spatial data using spatial statistics.

We will discuss ways of measuring the distribution of your spatial features, as well as ways that you can determine if your data has any spatial patterns.

Throughout the presentation, we will be discussing examples on how to use the ArcGIS Spatial Statistics tools. We will conduct some software demonstrations and we will have review periods followed by a question and answer session, during which time Lauren will answer some of your questions.

# Basics of spatial statistics

So let's get started with our first topic – the basics of spatial statistics and the tools available to you in the core ArcGIS 9 product.

## **What are spatial statistics?**

---

- **A measure of what's going on spatially**

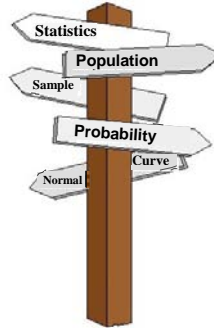
So, what are spatial statistics?

They are exploratory tools that help you measure spatial processes, spatial distributions, and spatial relationships.

There are a lot of different types of spatial statistics, but they are all designed to examine spatial patterns and processes.

## What are spatial statistics?

- A measure of what's going on spatially
  - Not the same as a-spatial statistics



Please understand that when we talk about spatial statistics we are not just talking about applying traditional, a-spatial or non-spatial statistics, like what you may have learned in your high school statistics class, to spatial data. When we talk about spatial statistics, we are describing specific methods that use distance, space, and spatial relationships as part of the math for their computations.

We're still going to talk about mean and standard deviation, but our focus will be in how those concepts apply to spatial features.

## What are spatial statistics?

- A measure of what's going on spatially
  - Not the same as a-spatial stats
- Two categories of spatial measurements

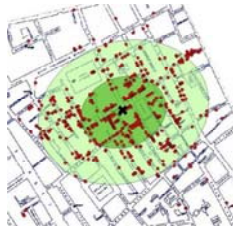


Today we will focus on a particular type of spatial statistics often referred to as pattern analysis.

The tools I'll be describing fall into two broad categories.

## What are spatial statistics?

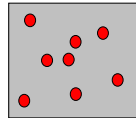
- A measure of what's going on spatially
  - Not the same as a-spatial stats
- Two categories of spatial measurements
  - 1) Identifying characteristics of a distribution



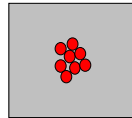
The first category is descriptive in nature. These tools quantify or identify characteristics of features. They answer questions like where is the center, or how are features distributed around the center?

## What are spatial statistics?

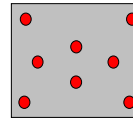
- A measure of what's going on spatially
  - Not the same as a-spatial stats
- Two categories of spatial measurements
  - 1) Identifying characteristics of a distribution
  - 2) Quantifying geographic pattern



**Random**



**Clustered**



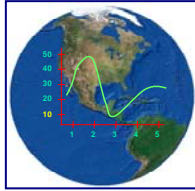
**Dispersed**

The second category is more concerned with describing spatial pattern. With these statistics, we are able to determine if our features are random, clustered, or evenly dispersed across our study area.

## Why use spatial statistics?

---

- To help assess patterns, trends, and relationships



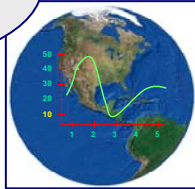
So why do we use spatial statistics?

Spatial statistics is one of many approaches that can help us to assess or analyze our data, and it's a pretty powerful too.

## Why use spatial statistics?

- To help assess patterns, trends, and relationships

Better understand  
behavior of geographic  
phenomena



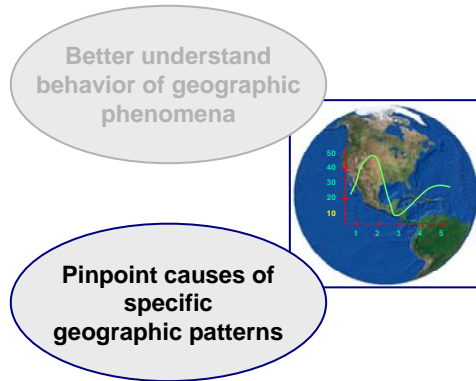
Why?

Because, with spatial statistics, we get a better understanding of geographic phenomena. If we think about it, everything happens in space and time.

When we analyze data outside of their spatial context, we really only get half of the story. As geographers and GIS analysts, we are especially interested in the spatial aspects and impacts of our data.

## Why use spatial statistics?

- To help assess patterns, trends, and relationships

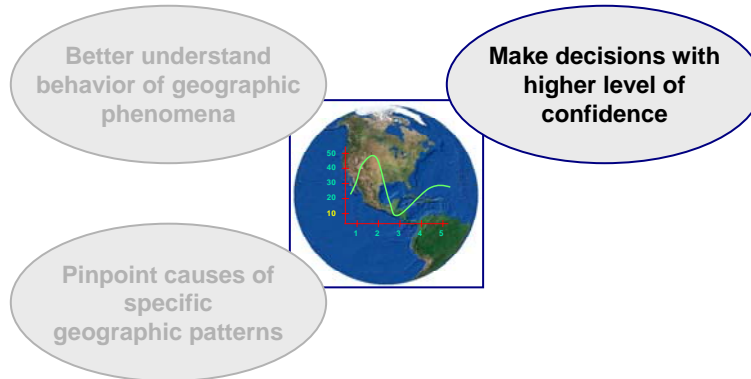


Spatial statistics can also help us to pinpoint causes of specific geographic patterns. One of the easiest analyses we can do is to map different phenomena and then overlay the maps to see if we notice correlations.

For example, we might find that a certain disease is only found in villages near rivers and streams. Perhaps our disease is being spread by a river parasite.

## Why use spatial statistics?

- To help assess patterns, trends, and relationships



Spatial statistics can also help us make decisions with a higher level of confidence.

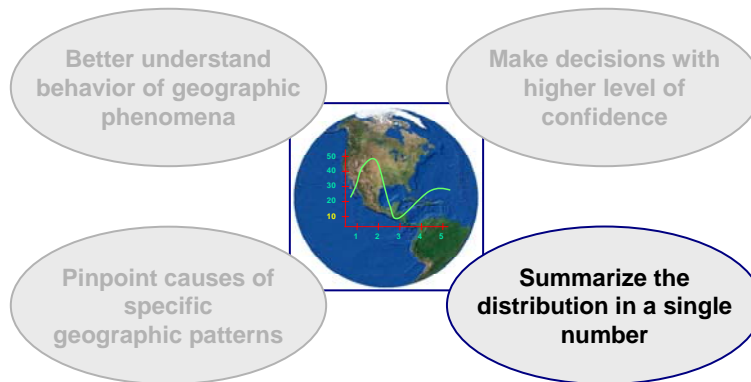
So you've done your analysis, and you've come up with some conclusions or recommendations.

Your boss is going to want to know how you came up with those conclusions, right?

You can say that “visual inspection indicates...” or, “my research suggests...” but it's great if you can also add “...and my results are statistically significant at the 0.05 level.” It gives you a lot more confidence in your decisions and conclusions.

## Why use spatial statistics?

- To help assess patterns, trends, and relationships

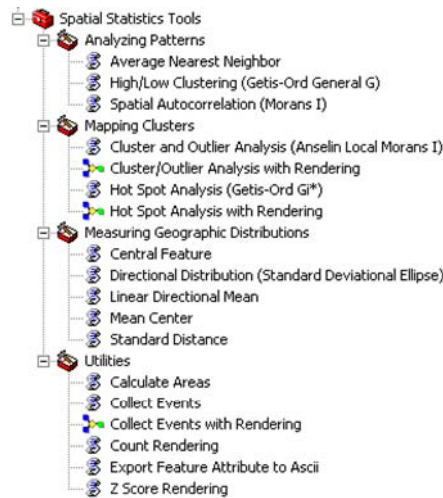


Finally, spatial statistics are helpful when we are dealing with large, complex data sets, which is exactly what we deal with in GIS everyday.

Spatial statistics often allow us to cut through some of the noise and complexity in our data to get straight at the broader trends or overall patterns. We are able to summarize our data and talk about averages, rates, or trends.

## Spatial statistics tools

- **Core functionality**
  - Not an extension
- **Available at all license levels**
- **Source code provided**
  - Scripts
  - Models



ArcGIS 9 comes packaged with spatial statistics tools needed for doing spatial distribution and pattern analysis. The tools are located in the Spatial Statistics tools toolbox, shown here. In this seminar, we will concentrate our time on three of the four toolsets.

It is important to understand that these tools are core functionality included in ArcGIS 9. They are not a separate extension that you have to purchase. Also, these tools are available at the all license levels: ArcView, ArcEditor, and ArcInfo.

What is really nice about these tools is the source code is provided. The tools are packaged in the form of scripts and models. You can look at the script and see the algorithms and mathematics for these tools, so they are great in a teaching environment. And, if needed, you can edit copies of the provided scripts to create new tools, or different implementations of these tools.

# Software demonstration

So, let's go to our first software demonstration.

All of our demos today will use crime data from a city in the United States, but I would like to take a moment to remind you that these tools can be used on a large variety of data, and for lots of different applications, such as epidemiology, archeology, and wildlife biology, just to name a few.

In this first demo, I want to emphasize why you might want to supplement traditional GIS analysis with spatial statistics.

Police resources are limited. They can't be everywhere all the time. So it's important to use the resources we have as effectively as possible. If we can identify where the crime hot spots are located, we can make better decisions about where we might want to increase law enforcement presence. Before we had GIS technology, police departments would typically use pushpin maps

to show spatial patterns in crimes. Let me turn on a digital replica of a pushpin map. So let's pretend that this is a wall map and that we've stuck pushpins in it to represent crime. In 2002, there were over 40,000 crimes reported in this city. Can we effectively determine our crime hot spots from this map? It looks like we need to have patrol cars almost everywhere. We really can't see more than a general city-wide pattern. So you can see that visually assessing the data with pushpins isn't very easy.

There was a natural progression in crime mapping as more police departments were using GIS and that was to create density maps of crimes. Now, to create a density map, I used the Point Density tool available with the Spatial Analyst extension. The result is a classified map that shows relatively where crime is high – those are the red spots, and where crime is low, that's the blue area. The tricky part about creating density maps is we have to make decisions about a couple of parameters. One of those is the neighborhood size. The distance value which tells the software how much area around each feature it should consider in the calculations. Selecting this distance can be subjective and the value you select can have a big impact on the results you get. The density we are looking at here was created using  $\frac{1}{8}$  of a mile. I also created a density map using a one-mile neighborhood. I used the same data; I just changed the distance value.

With that distance, I get a different picture, bigger concentrations and a lot more gradation. But this isn't the only choice we have to make with density maps. Once you have your values, you need to make a decision about how to render or symbolize the density values – which range of values will indicate high crimes, our red, and which range of values will indicate low crimes, our blue.

ArcGIS offers several options that you can choose from. In the next few maps, I'm going to show you, the underlying density values remained fixed. I use the results from the one-mile density

map that we're looking at here. The only thing changed is the rendering scheme. This map is rendered with the Jenks natural breaks. For this layer, I used equal area rendering. I can see that I get higher crime areas that are more compressed or compacted. Another option is quantile; with this, I get a much different picture. Crime is much more widespread. Again, we want to identify crime hot spots so we can make better decisions about how to allocate police resources, but we still haven't been completely successful with this. It's not that we can't see a pattern with the density maps; it's that we get different patterns depending on the distance value and rendering scheme that we use.

Here's where spatial statistics can step in and help out a great deal. This final map that I'm going to show you was created using the Hot Spot Analysis tool. Later, I will outline exactly how to do the Hot Spot Analysis. Right now, I want you to notice two things. First of all, unlike with the pushpin map where we had trouble identifying distinct patterns, here, it is easy to see where our hot spots are. They are shown in red. We can also see where we have cold spots – those places where we have very few crimes. The second thing I want to point out is that there is less subjectivity when we use spatial statistics than there is when we use density mapping.

The Spatial Statistics toolbox includes tools to identify an appropriate scale for my analysis and the rendering is straight forward. A census block is either statistically significant or it's not. So with a statistical approach to analyzing crime, we've accomplished our goal. We get a clear picture of crime hot and cold spots, which help us make better decisions about where and how to allocate police resources.

## Review and Q & A

---

- **What are spatial statistics?**
  - A measure of what's going on spatially
- **Why use spatial statistics?**
  - Quantify patterns and relationships
- **Spatial statistics tools**
  - Not an extension



In this section, we introduced spatial statistics as a method of exploring your data to see what's going on spatially.

We also talked about why you might want to use spatial statistics to quantify patterns and relationships in your data.

Finally, we talked about how the Spatial Statistics toolbox in ArcGIS 9 is not an extension but rather part of the core product available at all license levels.

I will now turn the seminar over to Lauren who will answer some of your questions.

Thanks Sandy. Our first question comes from the Woodland GIS class in Saskatchewan. They are asking, "Can you use the spatial statistics tools in Model Builder?"

The answer is yes. The spatial statistics tools work like all the Geoprocessing tools. You can use them in Model Builder. You can call them from the command line, you can run them as a tool dialog and you can also use them in scripts like Python scripts or VBA scripts.

Several people were asking questions about specific tools in Geostatistical Analyst and Spatial Analyst, so I want to tell you just a little bit about how the tools we're talking about today relate to the statistical tools in Geostatistical Analyst and Spatial Analyst. The tools that we're showing you today are core functionality and what that means is if you have ArcGIS 9, you have these tools. You also have the source codes for these tools.

The Geostatistical Analyst and Spatial Analyst products, those are extensions. That means if you want those tools, you would purchase them in addition to ArcGIS. The tools that we're showing you today are very different than the tools in the Geostatistical Analyst or Spatial Analyst.

Actually, there are several products that have other types of spatial statistics as well, including Business Analyst. We also have something called the SAS Bridge. But the statistics in those products are used for different applications and usually they're typically used to answer different types of questions. The Geostatistical Analyst product contains powerful tools that create surfaces or to predict values for a surface. So you would use Geostatistical Analyst if you were working with sample data. If you are in the mining or the petroleum industries for example, you're probably already using Geostatistical Analyst. The Spatial Analyst product includes very powerful tools for working with raster data. Now, raster data analysis has been around for a long time, so there's really a very rich collection of tools in the Spatial Analyst product. I'm including some Multivariate Statistical tools and also some very powerful tools to do hydrology-type analysis.

The tools in the Spatial Statistics toolbox that we're showing you today, we can say that they would fall into a category that we might call pattern analysis methods. In the academic literature, these methods are usually referred to as point pattern analysis. All the tools, with just one or two exceptions, work just fine with points, or lines, or polygon features. So it's really more appropriate to call them pattern analysis tools. I think with that, I'm going to turn back to Sandy.

Thank you Lauren.

# Measuring spatial distributions

Now let's talk about how we can measure the spatial distributions of our features.

## Measuring geographic distributions

- **Identify spatial characteristics of a distribution**
  - Where is the center?
  - What feature is the most central?
  - How are features dispersed around the center?

When we measure the distributions of spatial features we are interested in answering simple questions:

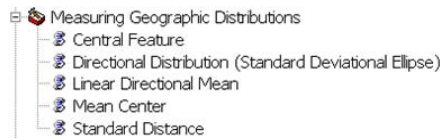
Where is the geographic center of all my features?

Which feature is the most centrally located?

How are features dispersed around the center?

## Measuring geographic distributions

- Identify spatial characteristics of a distribution
  - Where is the center?
  - What feature is the most central?
  - How are features dispersed around the center?
- **Measuring Geographic Distributions toolset**



The tools in the Measuring Geographic Distributions toolset are the tools you will need to answer these and other simple questions about your data.

## Measuring geographic distributions

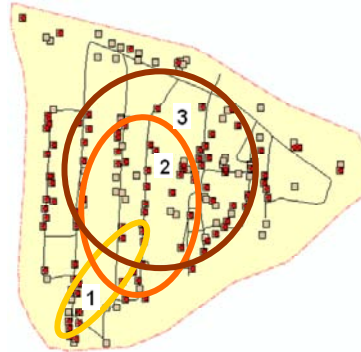
- **Identify spatial characteristics of a distribution**

- Where is the center?
- What feature is the most central?
- How are features dispersed around the center?

- **Measuring Geographic Distributions toolset**

- **Often used to:**

- Compare distributions of different features
- Identify directional trends of features
  - Examine changes over time



Often, these tools are used to compare the distribution of different features.

For example, in this graphic we are looking at cases of Dengue Fever for a village in India. We wanted to know how the distribution of the disease changed during the first three weeks after the outbreak. The Standard Deviational Ellipse tool, that we'll talk about in a minute, shows us that the disease started in the southwest corner of the village. It spread along the primary transportation corridors, and by week three, encompassed most of the village.

## Where is the center?

- **Mean Center tool**

- Computes the average X and Y coordinates of all features
- Creates a new point feature



So let's take a closer look at the tools in this toolbox.

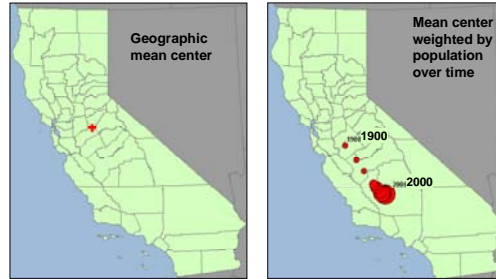
Now, when examining your feature distributions, one question you might ask is, where is the center of my distribution? To answer this question, you could use the Mean Center tool.

This tool will compute the average X and Y coordinate for all the features in your study area and generate a new point feature, indicating the center.

In the graphic, we have the counties in California. The mean center of all the California counties is located where you see the plus sign, just to the southwest of Yermo in California.

## Where is the center?

- **Mean Center tool**
  - Computes the average X and Y coordinates of all features
  - Creates a new point feature
- **Additional uses**
  - Compare distributions of different types of features
  - Track changes in the distribution



More commonly, we would use the Mean Center tool to compare distributions of different types of features, or to find the center of features based on an attribute value.

For example, using population data over the past 100 years as a weight for the Mean Center tool tells an interesting story.

In the early 1900s, the population center was up in the mid-section of the state, probably as a result of the gold rush. But, over time, the population center has shifted toward the southern end of the state, most likely in response to the manufacturing boom in aerospace engineering.

## What's the most central feature?

---

- **Central Feature tool**
  - Identifies the most centrally located feature
    - Feature having the lowest total distance to all other features

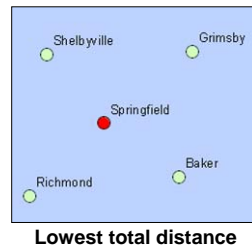
Another question you may ask about your data is, what feature is most centrally located? In contrast to the mean center, here we want to find the existing feature that is the most central to all other features. The tool we would use is called the Central Feature tool, and it identifies the feature that has the lowest total distance to all other features.

## What's the most central feature?

- **Central Feature tool**

- Identifies the most centrally located feature
- Feature having the lowest total distance to all other features

**What is the best location for a new health center?**



For example, let's say we are in charge of finding the best location for a new health center for these five cities.

In the graphic we see here, we don't really need a GIS to tell us which city is the most central. However, the Central Feature tool confirms what we see – that the most central city is Springfield.

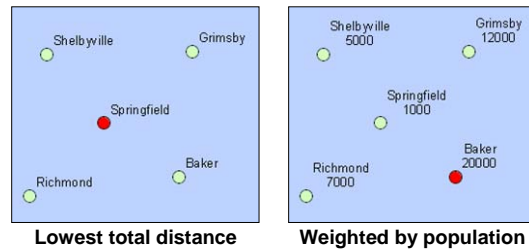
A more interesting example involves adding population and using that value as a weight in the analysis.

## What's the most central feature?

- **Central Feature tool**
  - Identifies the most centrally located feature
  - Feature having the lowest total distance to all other features
- **Additional use:**
  - Finding the most accessible existing feature

What is the best location for a new health center?

What is the most accessible location to the greatest number of people?

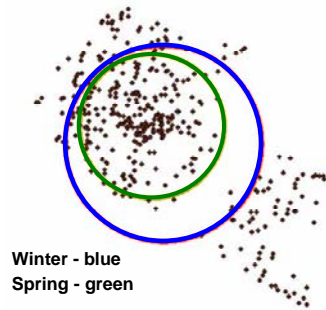


When we use population as a weight, we change our question a bit. We are no longer asking which city is most central, but which city is most accessible to the greatest number of people. Because Baker has a large population, placing the new health center there will maximize accessibility for the population as a whole.

## Measuring feature distribution

- **Standard Distance tool**
  - Measures distribution of features around the mean
  - Result is a summary statistic representing distance

Seasonal effects of childhood respiratory diseases



While the Mean Center and Central Feature tools tell us about the center of a distribution, they don't tell us about the overall distribution.

The Standard Distance tool tells us how dispersed our features are around that center.

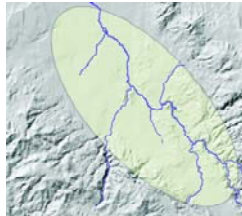
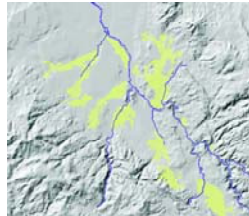
In the map shown, we have incidences for a respiratory disease among children. We want to see if we can detect any seasonality for this disease. For example, do we see a peak of incidents when certain plants or pollens are present?

To answer this type of question, we would examine the spatial distribution of the respiratory incidents by season. Output for the Standard Distance tool is a circle. When the circle is very large, we know that cases are widespread, like we see here for winter. When the circle is very small, we know that cases are more localized, like we see here for spring. Comparing the standard

distance circles for both seasons suggests this disease is exacerbated by the cold winter temperatures.

## Distributional trends

- **Directional Distribution (Standard Ellipse) tool**
  - Identify spatial trends in the distribution of features
- **Uses**
  - Compare distributions
  - Examine different time periods
  - Show compactness and orientation



In addition to learning about the distribution of features, you might also need to identify distributional trends. The Directional Distribution (Standard Ellipse) tool is similar in idea to the Standard Distance tool, but it allows you to create what are called standard ellipses indicating the orientation and direction of your distribution.

In our graphic, a wildlife biology study in bobcat movement between preferred habitat areas calculated the orientation of the movement areas to see if it coincided with natural features, such as valleys, rivers, or ridgelines. This could help provide clues to the travel routes the bobcats use.

# Software demonstration

Let's go to our second demonstration. In this demo, we're going to revisit our city from the first demo. In this demo, I want to illustrate some of the ways that we might use descriptive statistics that we just talked about with crime data. We will be answering two questions. The first one looks at crimes involving domestic violence and asks where should we locate a new support center to help victims of this particular type of crime. To get this dataset, I did a spatial join, which let me aggregate or count domestic violence incidences by census block for a particular community in our city. The community wants to create a new support center for victims of domestic violence. An obvious choice would be to put the new support facility right in the middle of the community.

If we run the Central Feature tool, we can find the most central census block, located, here, in the center of our community. But what if the community wanted to place the support center so that it better serves the population at risk? The map shows the number of domestic violence incidences

for each census block. To find the most accessible location to the population at risk, we would use the spatial distribution of past cases of domestic violence as a weight. Let's open the Central Feature tool. It's located in my Spatial Statistics toolbox in the Measuring Geographic Distributions toolset. For this tool, for my input, I'm going to use the domestic violence layer. That's the census blocks that we were just looking at. My output, I'm actually going to place in a geodatabase that I created just for these demos. I'm going to call it central feature domesticviolence\_w, so I've known I've used a weight with it. For my weight field, this is where I'm going to put the past occurrences and that column is the domestic violence column. I'm going to click okay to run the tool. So, it shouldn't take too long. Right now it's looking at our weights and seeing where the most central location is based on past occurrences of domestic violence. It looks like it's done. I'm just going to close ArcToolbox and I'm going to symbolize the results just a little bit differently so we can see them better. I'm going to left click on the symbol and I'm going to change the outline color to black and I'm going to increase the outline width.

So here we see our results and from it, the results, it appears that more cases of domestic violence are located in the western part of the community. If our goal is to make the new facility most accessible to the population at risk, this would be the good area to put it.

Now, let's use descriptive statistics to answer another question. We want to use prime attribute information to see if we can discern any temporal trends in occurrence. Our question is, does night time crime have a different spatial distribution than daytime crime, and we want to answer this question for each police beat in our city. If we observe a temporal difference in the crime patterns, then we will likely want to set up different patrol routes for the officers working the day shift, symbolized in blue, then for those working the night shift, symbolized in purple. We have constructed a unique code in this layer that identifies both the beat and the time that the crime occurred in. Was it in the day, or was it in the night? We're going to use this code as our case

field or our grouping field, for our tool. This is going to tell the tool to do a separate analysis for the day crimes and the night crimes in every beat.

I'm going to open ArcToolbox and, remember back to the lecture, we talked about two spatial statistics tools that help us look at the overall distributions of our features. They were the Standard Distance and Directional Distribution (Standard Ellipse) tools. The Standard Distance tool tells us how dispersed our features are around the mean center. The Directional Distribution (Standard Ellipse) tool tells us about the distribution of our features, as well as if there's a directional trend or orientation in that distribution. Because we want to know about the orientation of our crimes, we'll use the Directional Distribution tool.

Now, for my input, I'm going to use the crimes 2004, which are my point features of all my crime in census. My output ellipse I'm going to save in that same geodatabase and I'm just going to call it standard deviation ellipse. Now, in this situation, I want to use a case field. That's my grouping value that specifies the time and the beat in which my crime occurred. That column is this DorN\_beat column. Even though this analysis only takes about 40 seconds to run, I wanted to set up some symbolization to best communicate my results, so I've already run this analysis. I'm going to cancel here and just show you the results. I'm going to turn on the standard ellipse by beat layer that I created. My daytime crimes are symbolized as the light blue and my nighttime crimes are a dark purple. From the results, we can see that in most of the beats, there aren't any drastic differences between the occurrence time distributions.

Let's take a closer look. I'm going to zoom in to beat number six. As you can see here in beat six, the crimes for both the day shift and the night shift closely resemble a circle. The circular shape of the ellipse indicates there isn't a strong orientation to the occurrence of the crimes. So for this

particular beat, we can allocate police resources evenly throughout the beat during the day and night shifts.

Now, there is one beat where there appears to be a fairly strong difference between the day and the night and that's beat number nine. In this beat, the first thing we see is a difference in the size of our ellipses. The night crimes are much more concentrated than those during the day. We can definitely set up patrol routes accordingly, so that they are more concentrated during the night shift than the day shift. The second thing we see is the orientation of the crimes. Both ellipses have a southwest to northeast orientation. If we take a look at the major road network in this particular beat, we can see that there is a major road that runs from the southwest corner to the northeast corner of the beat. Could this road network be influencing the crime occurrence patterns? More analysis could help us determine if the roads are influencing crime locations.

## Review and Q & A

---

### ➤ Measuring geographic distributions

- Mean center
- Central feature
- Standard distance
- Directional distribution (standard ellipse)



In this section, we discussed ways of exploring the distributions of our features.

We answered the question, where is the mean center of my features, and which existing feature is the most central? The last two tools we talked about were Standard Distance and Directional Distribution.

I will now turn it back to Lauren, who will be again answering some of your questions.

Thanks Sandy.

We have a question from Mike in Rocky Hill. His question is, "Can we calculate a weighted mean center?"

Sandy showed us how we could calculate a weighted central feature, and yes you can. It works very similar to that. So for the Mean Center tool, you just specify the field that you want to use as

your weight. In fact, with ArcGIS 9.2, which is the next release, you'll be able to look at your features in three dimensions. So, it would calculate the mean x, y, and z.

We also got a call from Maria in Waggington. Her question is, "If your distribution is clustered into several clusters, could it be possible to get the center of each cluster, instead of the center of the whole distribution?"

The answer is yes. Many of the tools in the Measuring Geographic Distributions Toolset have a field called the case field. What the case field does is it allows you to group your features and to perform the analysis on each one of those groups. If you're asking if the tool automatically decides what the clusters are before hand, the answer is no. That would be a really good tool. At this point, you have to identify the features that are associated with each group and then run the tool. It'll compute the statistics separately for each group.

We also got a call from Andrea in Boise. She wants to know if it's possible to use more than one field as a weight when I'm calculating the mean center of the standard distance.

The tools that allow you to specify a weight really do only allow you to select a single field for that weight. Do keep in mind that you could always create a new column in your table and compute some type of an index based on more than one field that summarizes the weights of several fields into a single index and then use that index as your weight.

We got a call from Blake in Missoula. He asks about the Standard Distance and Standard Ellipse tools that Sandy talked about. They do notice that they ask the user to specify if they want one, two, or three standard deviations and what exactly does that refer to.

That question tells me that you're using these tools and that's great. Yes, both the Standard Distance and the Standard Ellipse tools ask the user to select the standard deviation level. The simplest explanation is that you get a larger circle or a larger ellipse when you choose a value larger than one. One is the default. The more detailed explanation is you have to think back to your high school statistics class here. You probably remember hearing about the normal distribution and the classic bell curve. Standard deviation is the measure of how the spread of your values are around the mean or around the average value in your dataset. In a spatial context, the standard deviation is a measure of how your features are spread around the mean center location. So when our values are normally distributed, we know that 68 percent of all the values will fall within +1 and -1 standard deviations of the mean. We know that 95 percent of our values will fall between +2 and -2 standard deviations of the mean. When we say we want to create a standard deviation ellipse with one standard deviation, we're asking for an ellipse that will encompass 68 percent of our features, if, and this is an important if, if our features are normally distributed around the mean. I'm going to turn it back over to Sandy.

Thank you Lauren.

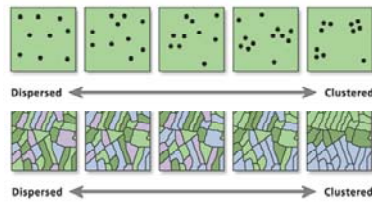
# Spatial pattern analysis

Now, let's talk about our next category of spatial statistic tools; spatial pattern analysis.

These tools give us ways to measure the degree to which our features are clustered, dispersed, or randomly distributed across the study area.

## Analyzing spatial patterns

- Using spatial statistics to determine feature distribution

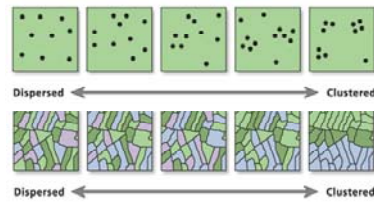


When we talk about analyzing spatial patterns, we are interested in finding out if there are underlying spatial processes influencing the locations of our features.

Are our features randomly located throughout the study area, or are they displaying a clustering or a dispersed pattern?

## Analyzing spatial patterns

- Using spatial statistics to determine feature distribution



- **Global calculations**

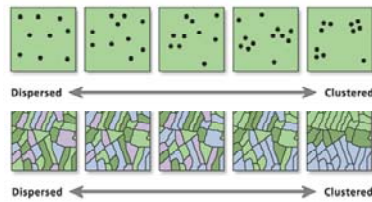
- Identifies the patterns/overall trends of data

We have two toolboxes in ArcGIS for analyzing spatial patterns. These toolboxes contain different approaches for analyzing patterns.

The first approach, global calculations, identifies the overall patterns or trends in the data. These types of statistics are very effective when we have a lot of complex messy data, and we are interested in understanding broad, overall trends. You can think of it, as if you were taking all your features and dumping them into the statistics crock pot. After they have simmered a bit, you will get a number (or two) that summarizes the overall spatial pattern of the data.

## Analyzing spatial patterns

- Using spatial statistics to determine feature distribution



- Global calculations

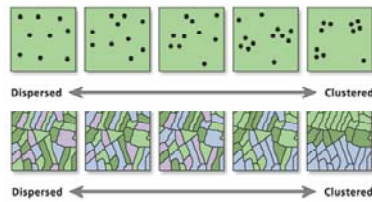
- Identifies the patterns/overall trends of data
- Analyzing Patterns Toolset

Global-type statistics are found in the Analyzing Patterns toolset.

They work by comparing your feature locations and/or attributes to a theoretical random distribution in order to determine if you have statistically significant clustering or dispersion.

## Analyzing spatial patterns

- Using spatial statistics to determine feature distribution



- **Global calculations**

- Identifies the patterns/overall trends of data
- Analyzing Patterns Toolset

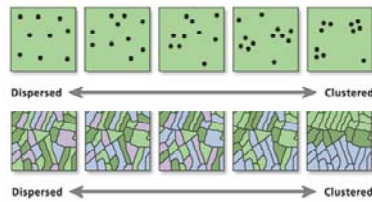
- **Local calculations**

- Identifies the extent and location of clustering or dispersion

The other type of statistics tools we have for analyzing patterns are categorized as Local Calculations. These calculations identify the extent and locations of clustering. They answer the question, where do we have spatial clustering?

## Analyzing spatial patterns

- Using spatial statistics to determine feature distribution



- **Global calculations**

- Identifies the patterns/overall trends of data
- Analyzing Patterns Toolset

- **Local calculations**

- Identifies the extent and location of clustering or dispersion
- Mapping Clusters Toolset

Local-type calculation tools are found in the Mapping Clusters toolset.

These statistics process every feature within the context of its neighboring features in order to determine whether or not it represents a spatial outlier, or if it is part of a statistically significant spatial cluster.

## Are features clustered?

---

- **Spatial Autocorrelation (Moran's I) tool**
  - Things that are closer are more alike than things that are not
  - Measures similarity of neighboring features
  - Identifies if features are clustered or dispersed

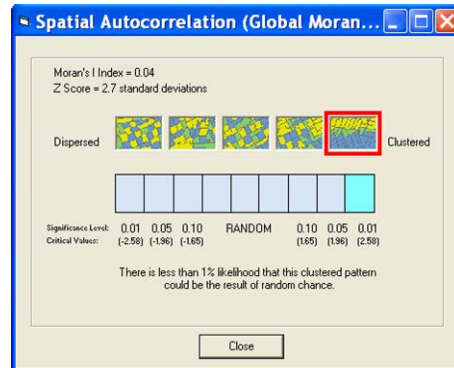
A global statistics question that we might ask about our data is, are features clustered?

The tool we could use to help us answer this question is called the Spatial Autocorrelation (Global Moran's I), found in the Analyzing Patterns toolset.

Okay, so what is spatial autocorrelation, you ask? It's a complex sounding term with a simple explanation. The concept is based on Tobler's First Law of Geography, which says everything is related to everything else – but nearby things are more related than things that are far away. So, with spatial autocorrelation, we are asking, are similar attribute values clustered in space?

## What is the overall pattern?

- **Spatial Autocorrelation (Moran's I) tool**
  - Things that are closer are more alike than things that are not
  - Measures similarity of neighboring features
  - Identifies if features are clustered or dispersed



The output for this tool is a graphical display that gives you the results 4 different ways.

It starts off at the top, giving you the statistical numbers.

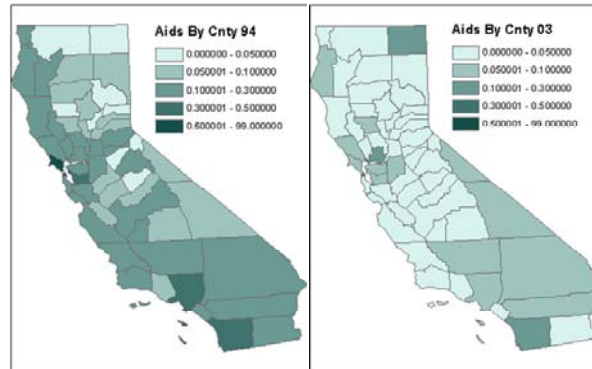
Under that, we have a pictorial representation of what those statistical numbers mean. The picture tells you if the pattern is dispersed or clustered.

Under the graphic, we have a bar graph that shows if the results are statistically significant.

Finally, at the bottom of the dialog, the results are presented as a sentence. For the analysis presented here, the sentence says, “There is less than 1 % likelihood that this clustered pattern could be the result of random chance.”

## Are features clustered?

- **Spatial Autocorrelation (Moran's I) tool**
  - Things that are closer are more alike than things that are not
  - Measures similarity of neighboring features
  - Identifies if features are clustered or dispersed

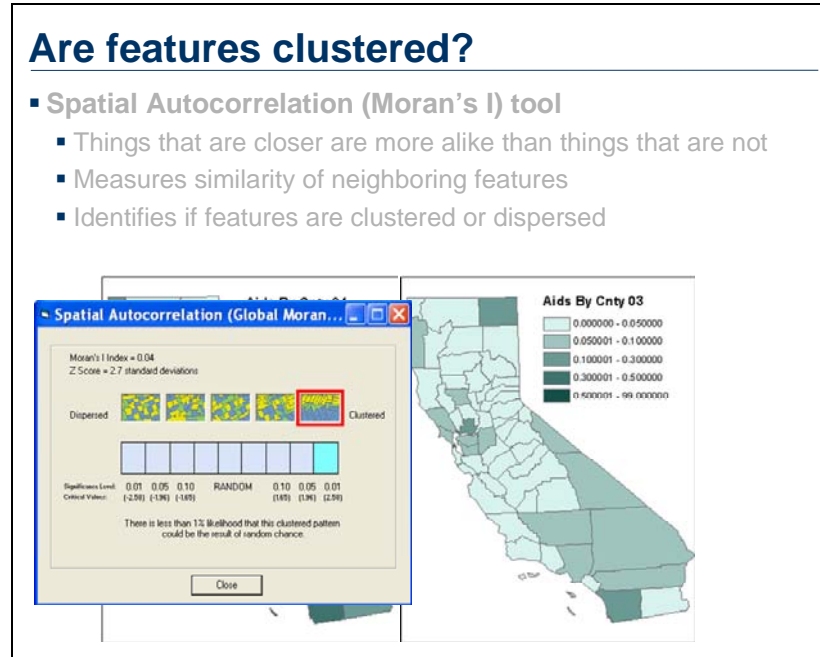


Let's look at an example. Before I explain it, please know that the analysis and results being presented here were done for demonstration purposes only. If we wanted to confirm or to publish these results, we would have to do a more thorough investigation.

In any case, we are showing here the number of new AIDS cases by county in California for both 1994, here on the left, and 2003, here on the right. Our question is, how is AIDS spreading? Is it rapidly spreading outward from the high risk areas to surrounding the counties, or is it remaining geographically fixed?

Knowing this could help health authorities determine how to best address the problem. In the two thematic maps showing AIDS cases by county, we don't really get clear understanding of how the disease is spreading:

Overall we know that the total number of new AIDS cases decreased each year between 1994 and 2003 – and we see here that the distribution of new cases for each county is changing, but how?



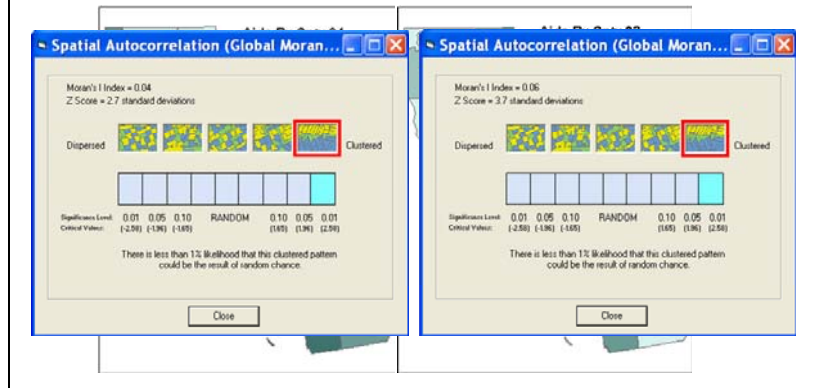
Let's look at the results of the Spatial Autocorrelation analysis. For 1994, we can see that new AIDS cases are clustered.

Counties with a lot of new cases are found next to counties that also have a high number of new cases.

Now we ask, do they remain clustered in 2003? Is clustering of new AIDS cases becoming more intense, or less intense, over time?

## Are features clustered?

- **Spatial Autocorrelation (Moran's I) tool**
  - Things that are closer are more alike than things that are not
  - Measures similarity of neighboring features
  - Identifies if features are clustered or dispersed



When we look at new AIDS cases for 2003, we see that the Z score has increased. In 1994, it was 2.7 standard Deviations, and in 2003, it is 3.7 standard deviations.

This means that the clustering is more intense than it was in 1994. When all the years were processed from 1994 to 2003, the overall pattern was a slow increase in clustering, indicating new cases are indeed remaining geographically fixed.

## Locate the hot spots

---

- **Hot Spot Analysis (Getis-Ord  $G_i^*$ ) tool**
- **Indicates the extent to which each feature is surrounded by similarly high or low values**

In the previous slide we asked a global question – is there clustering?

Now we are going to switch gears and ask a local question, where are the clusters, or where do features with similar attribute values cluster spatially together?

One of the tools in the Mapping Clusters toolset is called the Hot Spot Analysis Getis Ord  $G_i^*$  statistic, and it can be used to delineate clusters of features with values significantly higher or lower than the overall study areas mean or average value.

## Locate the hot spots

---

- Hot Spot Analysis (Getis-Ord  $G_i^*$ ) tool
- Indicates the extent to which each feature is surrounded by similarly high or low values
  - Identifies where clustering occurs in both high and low values
  - Calculates a Z score for each feature
    - High Z = hot spot
    - Low Z = cold spot

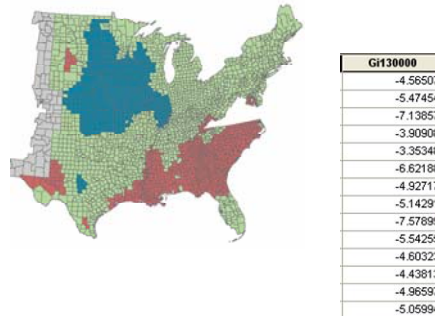
This tool identifies clustering in both the high and the low attribute values. A standardized Z score is calculated for each feature.

A high Z score results when a feature has a high value and it is surrounded by other features with high values. This is a hot spot.

Similarly, a low Z score results when we have features with low values surrounded by other features with low values. This is a cold spot.

## Locate the hot spots

- Hot Spot Analysis (Getis-Ord  $G_i^*$ ) tool
- Indicates the extent to which each feature is surrounded by similarly high or low values
  - Identifies where clustering occurs in both high and low values
  - Calculates a Z score for each feature
    - High Z = hot spot
    - Low Z = cold spot
- Useful in:
  - Crime prevention
  - Locating target markets
  - Finding the source of an epidemic



An example we have for Hot Spot Analysis uses mortality data for the eastern half of the United States.

We wanted to know if there are persistent areas in the United States where people are either dying earlier, or living longer, than the average American.

We used Hot Spot Analysis to analyze 20 years worth of mortality data. The counties shown here in red are those that were statistically significant early death hot spots for all 20 years.

Similarly, the counties shown in blue are those that are statistically significant cold spots for early death for all 20 years. In other words, folks in the counties shaded blue, persistently live longer than the average American.

# Software demonstration

Let's continue with our crime analysis. In the first demo, I showed you a hot spot analysis map for our city, based on all crimes, and I promised I would give you more details about how to do Hot Spot Analysis, so that's what we're going to do now. First, the context of the analysis.

One of the communities in our city came to us and asked us to help them locate a suitable area for a vandalism prevention program. Their goal is to build a facility that will help reduce the spread of vandalism. One possibility would be to use the Hot Spot Analysis tool to find those neighborhoods with lots of vandalism. We are looking at the vandalism per census block right now. But you can probably guess where the neighborhoods with a lot of vandalism are going to be. We expect to find lots of vandalism in the same neighborhoods where we have lots of crime, typically in our downtown and high density neighborhoods.

Let's see. I have a map here of our total crime per census block. As we can see, we have a very similar picture. Our high vandalism is in fact in the same areas as high crime. An alternative is to

locate the prevention program in neighborhoods where vandalism is higher in proportion to total crime. To analyze this, we would divide the number of vandalism incidences by the number of all crime, creating a ratio of each census block. If we look at the ratio of vandalism to all crimes, we see a different pattern. It seems to be a bit more scattered and probably less concentrated in the city center. It's pretty interesting, but let's keep going and see what our statistics tell us.

Now, performing a crime Hot Spot Analysis involves three steps: aggregating the incident data, determining the most appropriate scale of the analysis, and then running the Hot Spot tool. The tools we will be using require weights, so our first step is to aggregate our crime data. I've already done that aggregation using a spatial join and we looked at the counts of all crime and the counts of our vandalism already. The second thing that we need to do, before we run our Hot Spot Analysis tool, is to find an appropriate scale for our analysis. To do this, we use the Spatial Autocorrelation, Global Moran's I tool.

The Spatial Autocorrelation tool is in the Analyzing Patterns toolset. So I'm going to open ArcToolbox and I'm going to go to that toolset. This tool is going to help me determine the distance for the spatial processes promoting spatial clustering are most pronounced. Our input feature class for this tool is going to be our census blocks and I'm going to use the Count All Crime per block. The input field is the attribute value that contains my weight or my count. We're interested in finding out where we have clustering of vandalism that is higher in relation to all crimes. So to find the most appropriate distance for evaluating clustering, I'm going to use the count of all my crimes, TotalCrime02. This way, the distance that I find can act as a baseline if I want to run the same analysis for crimes other than vandalism. I'm going to check this box for displaying the output graphically.

Now, conceptualization is how we define the relationship among our spatial features. We have several choices here, including a user defined spatial weights matrix. The recommended conceptualization for the Hot Spot Analysis tool is the Fixed Distance Band. So that's what I'm going to use. For our Distance Method, I'm going to use Euclidian, which uses a straight line distance in the calculation. The last parameter that I'm going to fill out is the distance band. That is the distance that defines the range of spatial interaction among my features. We, again, want to know the distance where spatial autocorrelation is maximized. So to do this, we're going to run the tool multiple times. We'll start using a distance of a quarter mile, 1,320 feet. Then, we're going to increase the distance until we find a peak in spatial clustering. I'm going to click OK to run the tool and it's just going to take a few seconds.

Now while that's running, I want to let you know that the best way to examine how spatial clustering changes as the scale of analysis changes, is in fact to run the Spatial Autocorrelation tool multiple times. In the release of ArcGIS 9.2, there will be a new tool called the Multi-Distance Spatial Cluster Analysis tool, based on Ripley's key function that will provide a more straightforward way of finding this value. So it looks like we're just about done processing. It's calculating our results for us and it's going to open for us the graphic that we saw earlier in the slides, the one that shows us about our clustering and our dispersion of our features.

Okay, so here we have our results. Let's take a look. We definitely have clustering and our Z score is 15.4 standard deviations. That'll be important in just a few minutes and it's significant at the 0.01 level. This is good. So now we know that our crimes are clustered and we know a distance where clustering is significant. But we don't yet know what distance clustering reaches a maximum or peaks. I'll show you that in just a minute. Right now, I want to start the Hot Spot Analysis tool so it can run while we talk about that distance measure.

To locate our hot spots, we're going to use the Hot Spot Analysis with Rendering tool, located in the Mapping Clusters toolset. Not only are my stats going to be computed for me, but each feature is going to be symbolized as well. So my input feature class is again going to be my census blocks. My input field is going to be the ratio column, where I calculated the ratio of vandalism to total crime. Because this tool is going to do a symbolization for me, it wants to create a layer file. I have a layer file directory, where I'm actually going to save my layer, and I'm going to call it RatioHotSpot3960.lyr. The 3,960 is the distance that I'm going to use and I just personally like to put that as part of the name. Now, my output feature class is going to be a new dataset with my Z scores computed for every feature. So I'm going to put that in my results geodatabase. I'm going to call it the same name but I'm going to get rid of that .lyr. Now, my Distance Band that I'm going to use is 3,960 feet. That's the distance where crime clustering is the most significant. I'll show you how I know that in just a moment.

So we're done with this dialogue. Let's get it running and talk about that distance then. So how do I know to use  $\frac{3}{4}$  of a mile? I had actually run the Spatial Autocorrelation tool multiple many times, each time increasing the distance band. Here's a graph of my results. We started with a distance of a  $\frac{1}{4}$  mile. We saw that we had clustering of our crimes and our Z score was 15.4. When I ran the spatial autocorrelation tool for  $\frac{1}{2}$  mile, my Z score increased to 25.9. At  $\frac{3}{4}$  of a mile, it was still increasing, this time to 29.6 and at one mile, I saw it start to decrease. It's now 29.3. The distance where my Z score is highest is where the spatial processes promoting clustering is most pronounced. That's why I used the  $\frac{3}{4}$  of a mile, 3,960 feet, in our hot spot analysis. Speaking of which, it looks like our analysis is done.

Let's take a look at the results. Remember, we didn't want to place our prevention program where vandalism was simply the highest, but rather where it was a higher than expected proportion to all

other crimes. From our hot spot analysis, we can see that vandalism is a larger proportion crime in the suburban areas of our community. If we looked at the raw counts of vandalism, we might conclude that vandalism is an urban or downtown issue. But that's not what we see when we use the proportion. The areas shown in red are the areas where the ratio of our vandalism to all crime is highest. The blue areas are our cold spots and that's where the ratio is lowest. We now have recommendations for our community's vandalism prevention program.

## Review and Q & A

### ➤ Analyzing patterns

- What is the overall pattern?
  - Global calculations
  - Analyzing Patterns toolset
- Where are the clusters?
  - Local calculations
  - Mapping Clusters toolset



In this section, we covered how to analyze patterns in spatial data. We discussed tools that assess the overall pattern and we talked about tools that help identify locations of spatial clusters.

I'm going to turn it over to Lauren one last time to answer a few more of your questions.

Thanks Sandy. We got a question from Yuri in Delta, and the question is, "Can I add directionality when calculating the global Moran's I statistic, the Spatial Autocorrelation tool?"

The answer is yes. For all of the spatial autocorrelation type-tools, there's a parameter that's called your conceptualization of spatial relationships, and some of the options that you're given are inverse distance, which means you see relationships as things that are close by are more important than things that are farther away, so it's kind of a distance decay. Another option is fixed distance band. This models spatial relationships in terms of a sphere of influence. The last option is to get

your relationship values from a spatial weights matrix file. So, you could actually model time, and you could model any kind of directionality that you wanted. What you do is you create a file that contains the weight from and to every single feature, or the features that have a weight, and you would use that weights file when you ran these statistics.

We also have a question from Ivy in Atlanta. She asks, "To use these measures, you have to aggregate them into polygons, right? You can't just check spatial autocorrelation or use these tools with just the point locations?"

The answer to that question is that all of the tools work fine with points, lines, or polygons. But what I think she's getting at is that a lot of times we want to use these tools, Hot Spot Analysis for data that's incident-type data, like crimes or disease. That often comes in as one record per incident and there's no weight. In that case, we can still run these tools, but these tools do need a weight.

One way to get a weight is certainly, as she says, is to aggregate your crime counts; for example in the census block, or to recreate quadrates, or square polygons and count the number of crimes in your quadrate and use it that way. If you have very dense data, and you end up where at least some of your points are coincident, they have the same x and y coordinates, you can actually use a tool in the Utilities toolbox for the spatial statistics tool called Collect Events. What Collect Events does is it looks at all your point data, say it's incident data, and it identifies any data that's coincident, that has exactly the same x and y coordinates, and creates a new output. The output is a point feature class with a weight. The weight is the number incidents that occurred at that location. A lot of people ask us if we have Fuzzy Collective tool that says that these two points are really close, they're not exactly the same x and y coordinates, how do I add a fuzzy tolerance to the Collect Events tool? There's a good solution to this actually. If you run the Integrate tool,

also in the Geoprocessing toolbox, it's called the Integrate tool, before you run the Collect Events tool, what you'll be able to do is enter a fuzzy tolerance and snap points that are very close together to the same x and y coordinates, and then get your weighted point data from the Collect Events tool.

Ron from Cedar Falls is asking, "What is a Z score?"

A Z score is simply a measure of standard deviation. So, for example, if the statistic that you are running, Spatial Autocorrelation or Hot Spot Analysis, if it returns a Z score of 2.5, you would interpret that as 2.5 standard deviations. A Z score is a reference value that's associated with a standard normal distribution. So a very, very high or a very, very low Z score would be found in the tails of the normal distributions. For these Pattern Analysis tools, when you get a very high or a very low Z score, what you're finding is that you have a pattern that deviates significantly from a hypothetical random pattern. So to give an example, the critical values for Z scores when you're using a 95 percent confidence interval are -1.96 and +1.96 standard deviations. So if your Z score is between -1.96 and +1.96, you cannot reject your null hypothesis. You're seeing a pattern that could very likely be one version of a random pattern. But when your Z score falls outside that range, if you get a very high or very low score, like -2.5 or +5.4, you have a pattern that's too unusual to be a pattern of random chance. So we can reject the null hypothesis, so we can figure out what spatial processes might be causing that pattern either or a clustered or dispersed pattern.

I'm going to turn it back over to Sandy.

Thank you Lauren.

## For more information

---

### **Instructor – led training**

*Advanced Analysis with ArcGIS*  
*Introduction to Geoprocessing Scripts using Python*



### **Other resources**

*ESRI Guide to GIS Analysis, Volume 2*  
Andy Mitchell ESRI Press

*Using Spatial Statistics Tools demo*

Location suitability for 911 emergency response centers  
([www.esri.com/software/arcgis/arcinfo/about/demos.html](http://www.esri.com/software/arcgis/arcinfo/about/demos.html))

Now, before we say good-bye, I'd like to point out some resources available to you.

We have a couple of instructor led classes that will help. We have the Advanced Analysis with ArcGIS class that has a full chapter in the Spatial Statistics tools and we have the Introduction to Geoprocessing Scripts with Python that will help if you need to actually get into the Python scripting of the tools.

Another resource we have is the ESRI Guide to GIS Analysis Volume 2, by Andy Mitchell from ESRI Press. This is a great resource for learning about these tools. Every chapter goes through the different tool sets in the Spatial Statistics toolbox.

In a few weeks, the recordings of this seminar will be available for free on the ESRI Virtual Campus. The resources listed on this slide and more will be accessible from the recorded seminar. We hope you enjoyed today's seminar and on behalf of ESRI, I'd like to thank you all for attending.

