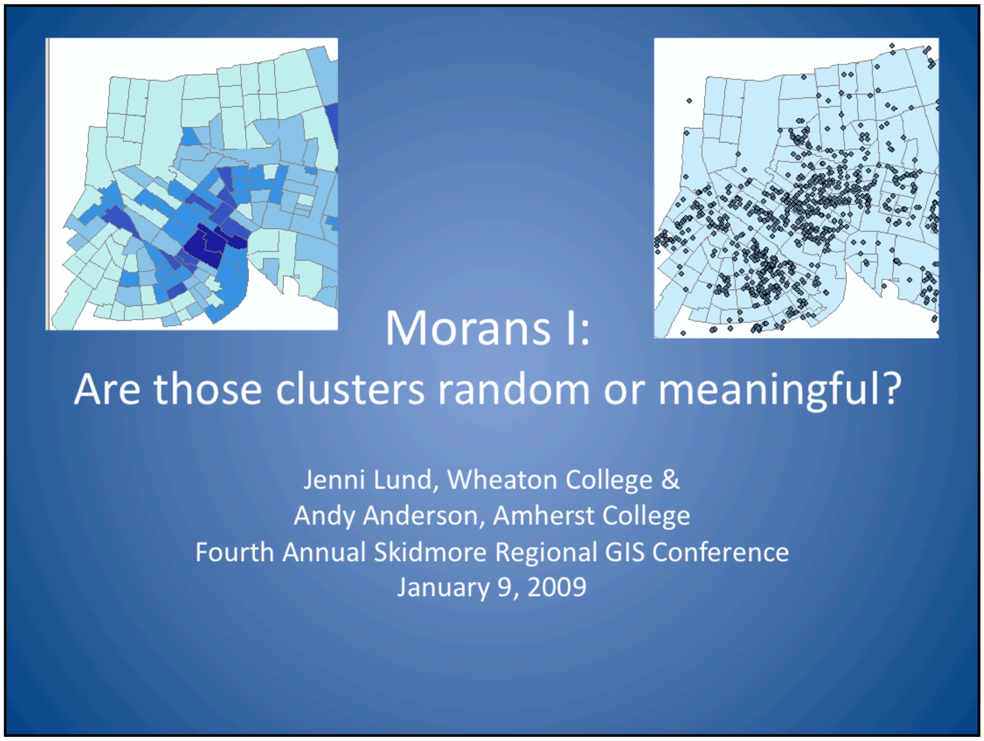


Morans I:
Are those clusters random or meaningful?

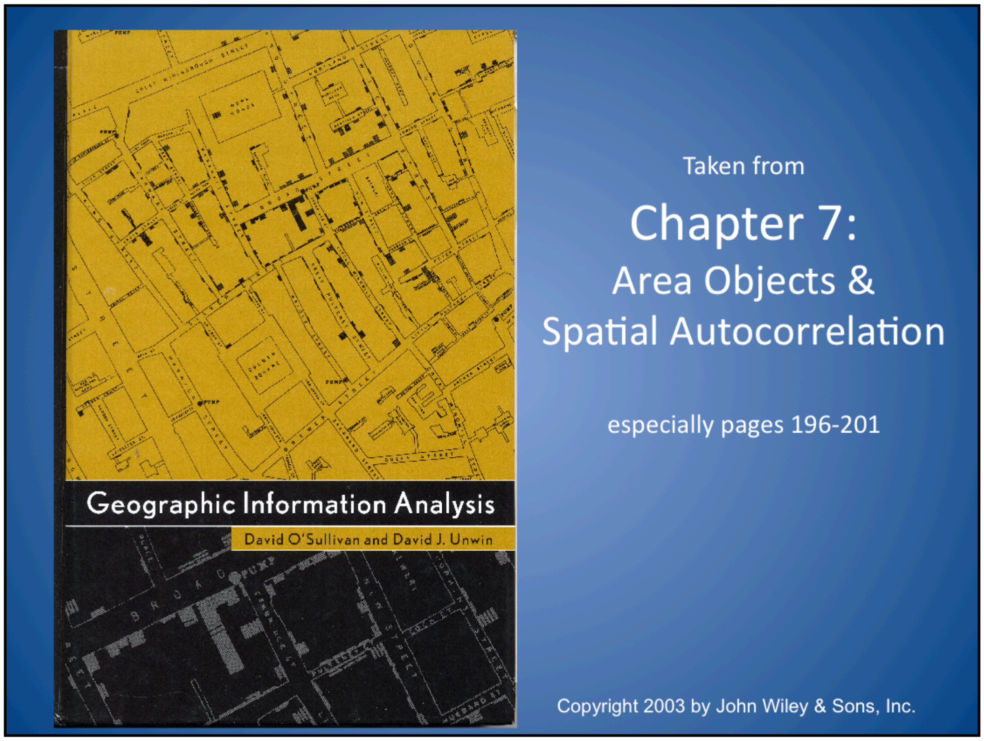
Jenni Lund, Wheaton College &
Andy Anderson, Amherst College
Fourth Annual Skidmore Regional GIS Conference
January 9, 2009

So far in my GIS career, I've used what I call "eyeball analysis" --- I make maps and say, "Golly. There seem to be a lot more over there than over here." I know enough about statistics to know that not everything that looks like a cluster really IS a meaningful cluster. It could be just a cosmic accident, like those shapes that show up when you're knitting with variegated yarn. I'm going to show you a way to check whether or not it's random or meaningful.

I confess that I usually see a meaningful cluster when I can think of a reason for the cluster. I might say, "Yes, that's a meaningful cluster of crime in the center of the city, because everyone knows that cities are full of crime." However, I've been wishing for a bit more rigor.

Andy Anderson and I were both interested in bringing more statistical analysis to our mapping, so we started reading a book and calling each other to talk about it: an ad hoc study group like the one that Sharron described. Andy actually understands the math, and I'm experiencing a gradual illumination with his help.

He is the one who actually figured this out, but he couldn't be here today. He sends his regards,


Geographic Information Analysis
David O'Sullivan and David J. Unwin

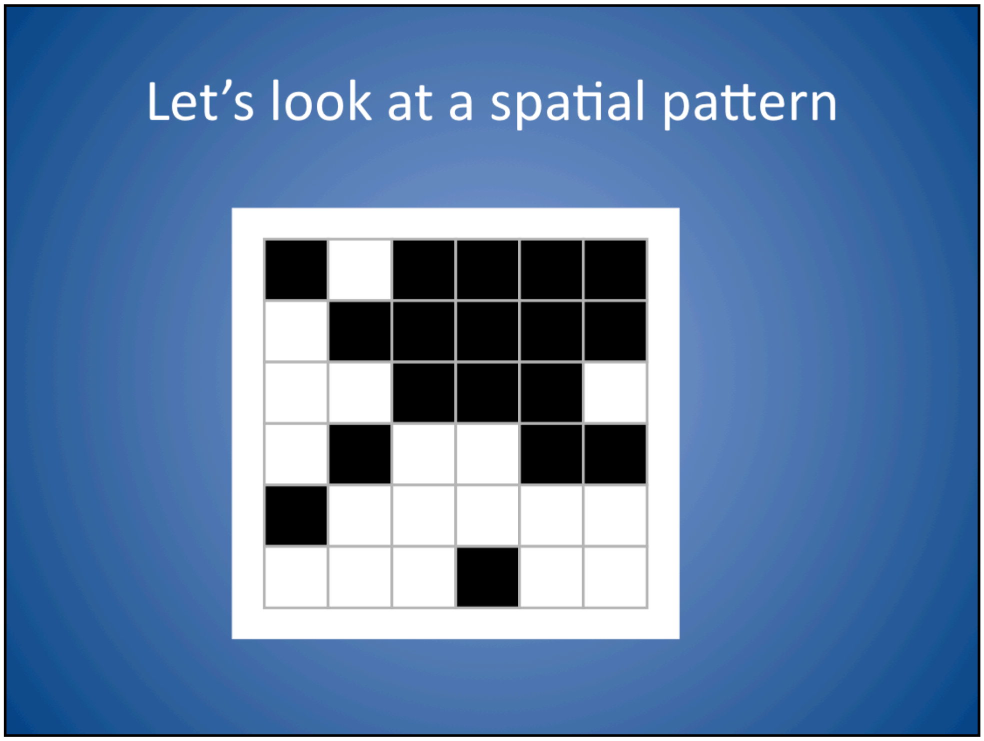
Taken from

Chapter 7:
Area Objects & Spatial Autocorrelation

especially pages 196-201

This is the book we used. In 2007, Diana Sinton organized a NITLE workshop at Wheaton on Spatial Statistics. Dave Unwin teleconferenced in as our instructor, and Diana and Bill Huber from U.Penn were our lab instructors.

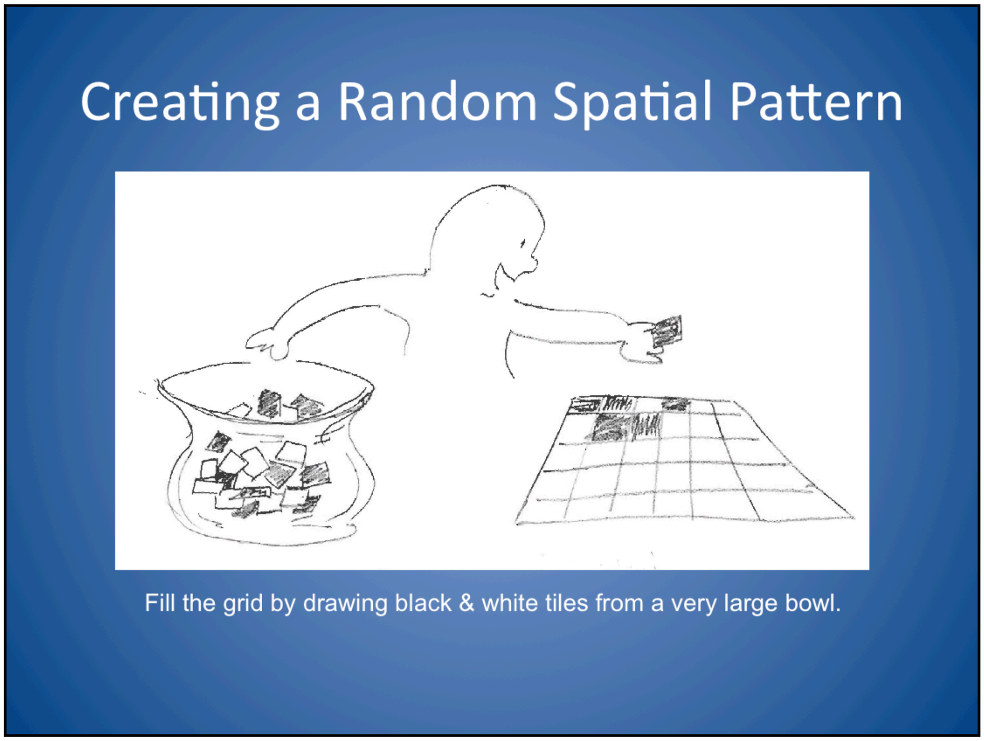
Finally, last semester, I had an impetus to use a spatial statistic. A student in our GIS course was mapping crimes in New Orleans, and he wanted to map hotspots. This is one tool Andy applied to the question, to help me out, and this is the one I feel confident about explaining.



Let’s look at a spatial pattern

■	□	■	■	■	■
□	■	■	■	■	■
□	□	■	■	■	□
□	■	□	□	■	■
■	□	□	□	□	□
□	□	□	■	□	□

When we look at our choropleth maps, we’re looking at patterns in space. To get warmed up to the subject, let’s look at a simple grid like this.



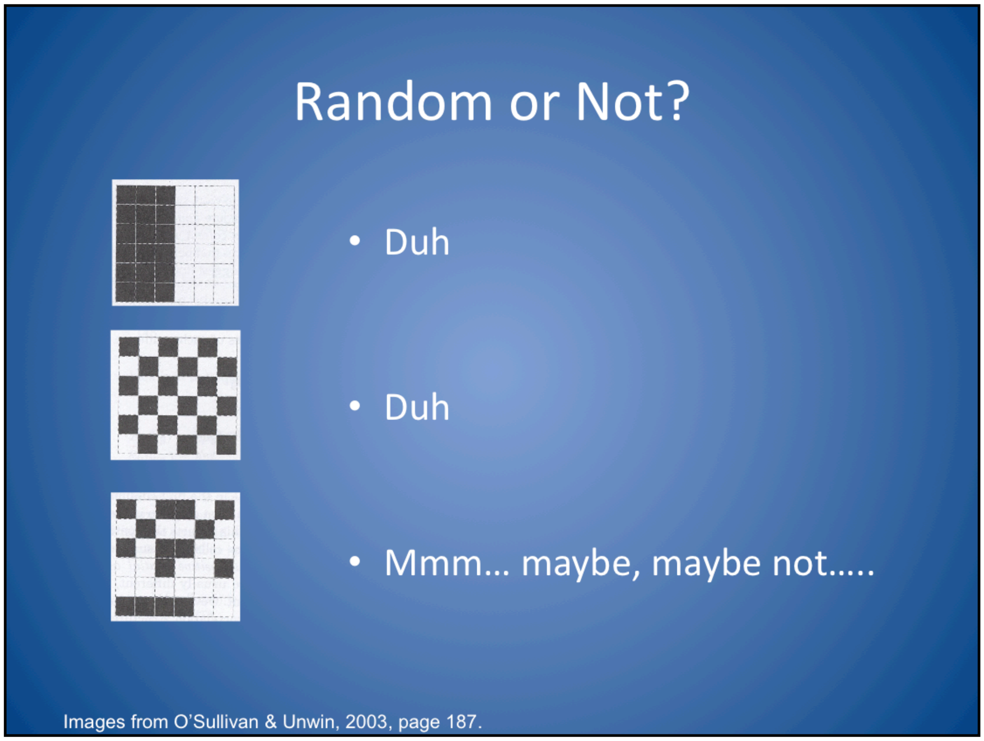
Creating a Random Spatial Pattern

Fill the grid by drawing black & white tiles from a very large bowl.

If we were going to create a grid with a random pattern, we could do it this way:

we'd reach into a huge bowl full of black and white tiles,
we'd pick out a black or a white tile, and lay it in the first square of the grid.
We'd fill the whole grid that way, by picking a black, then maybe a white and maybe another white and then maybe a black, until we're done.

Then we'd have a random spatial pattern.



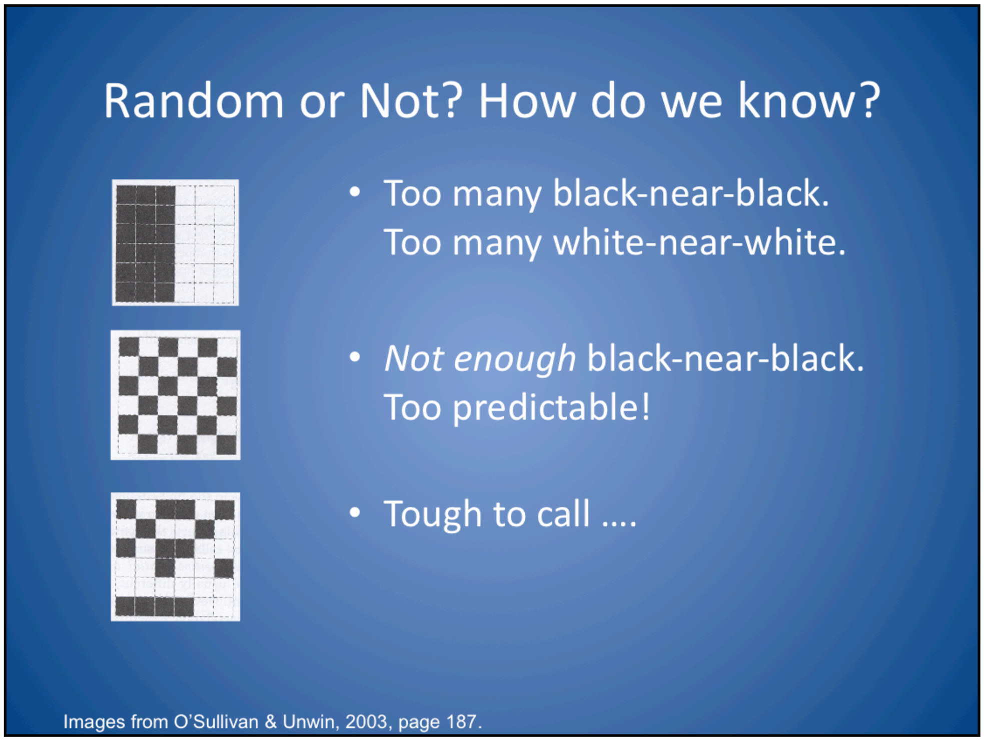
Random or Not?

- Duh
- Duh
- Mmm… maybe, maybe not…..

Images from O'Sullivan & Unwin, 2003, page 187.

Here are some possible patterns.

How likely is it that we're going to end up with a spatial pattern like this? Is this random?



Random or Not? How do we know?

- Too many black-near-black.
  Too many white-near-white.
- *Not enough* black-near-black.
  Too predictable!
- Tough to call ….

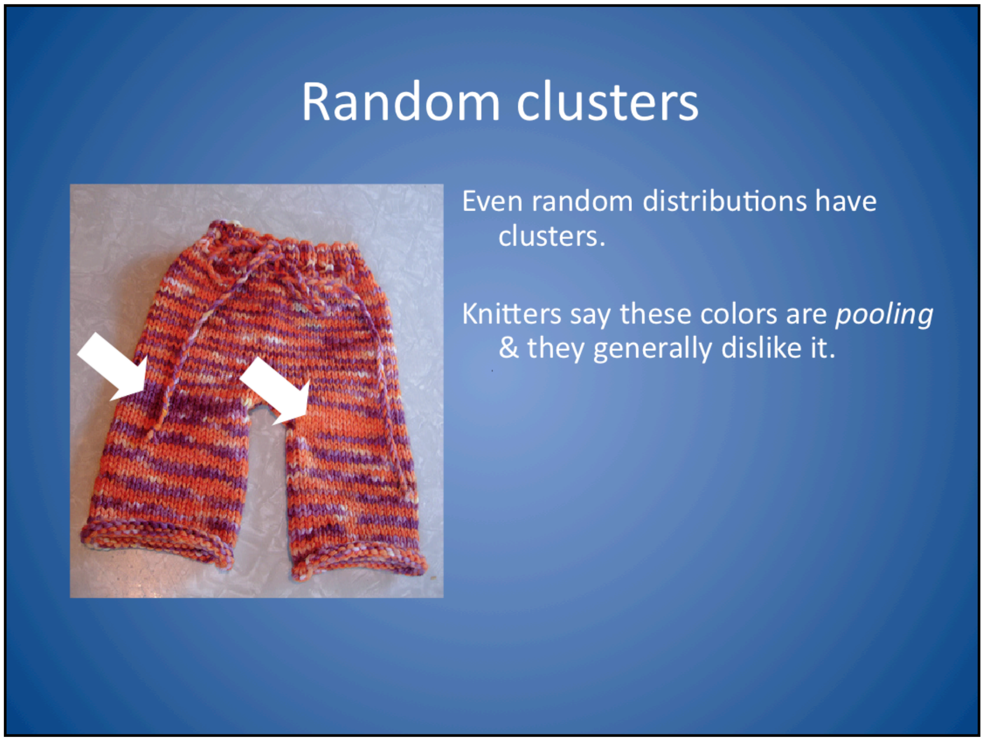
Images from O'Sullivan & Unwin, 2003, page 187.

We're sure the first one isn't random. Why? How do you know that?
You did an intuitive "test" in your head, and the results said, "There are too many black-near-black and white-near-white. That would never happen by accident."

Well, it MIGHT happen by accident … if you picked tiles for a gazillion years, this might actually happen once.

The second one isn't random, either. Why? How do you know that?
Because when you did that "test" in your head, and you said, "There are too FEW black-near-black and white-near-white. That would never happen by accident."

When you see a predictable pattern like in the top two, you know it is NOT random. But what about the third one? We need to have a way to measure whether that would have happened by accident. The nature of randomness is that there will be clusters, like with the variegated yarn.



Random clusters

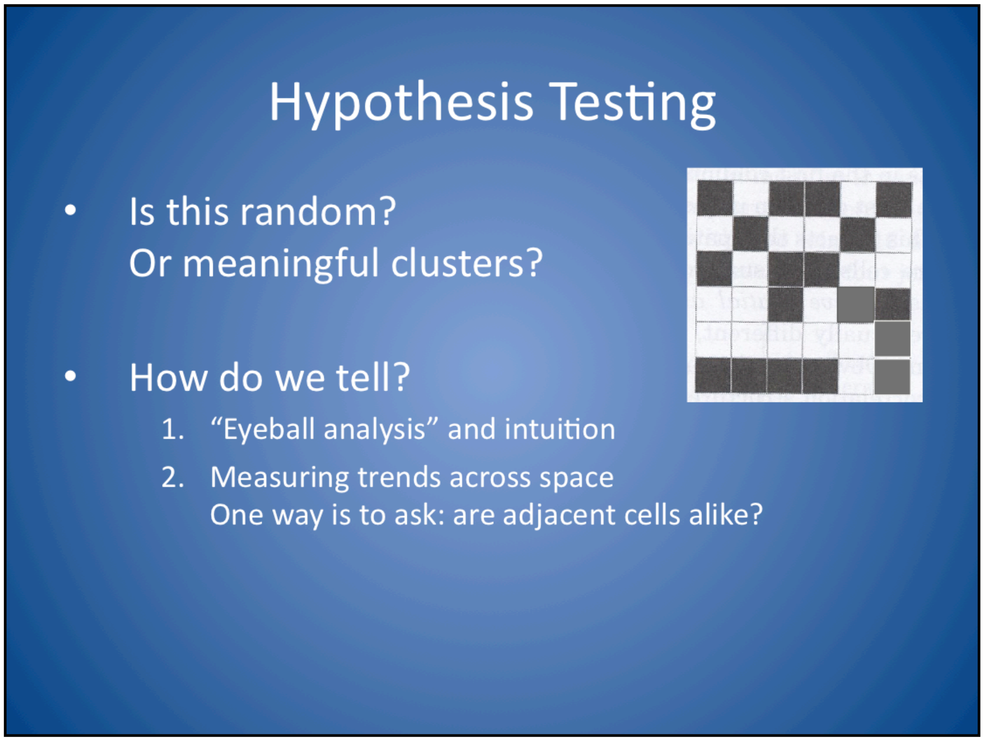
Even random distributions have clusters.

Knitters say these colors are *pooling* & they generally dislike it.

Variegated yarn is a single strand dyed in alternating colors.

The colors are distributed randomly. Distribution depends on many things (length of each row, size of the knitting needle, etc.)

Even random distributions have clumps. Knitters call it “pooling” when clumps of colors appeal.



Hypothesis Testing

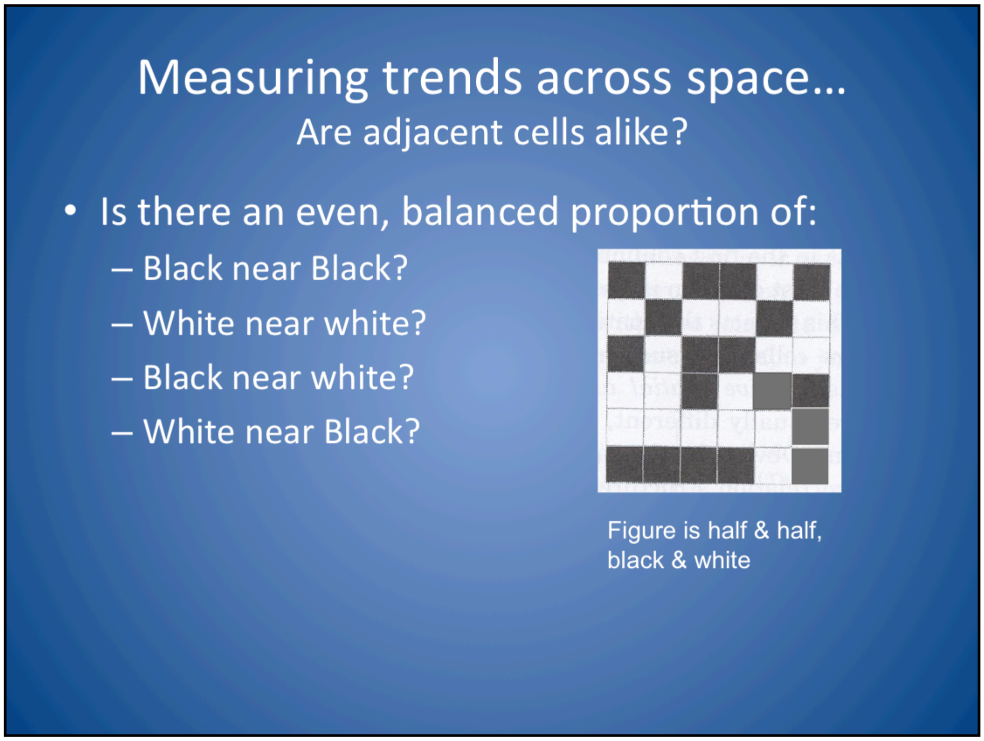
- Is this random?
  Or meaningful clusters?
- How do we tell?
  1. “Eyeball analysis” and intuition
  2. Measuring trends across space
     One way is to ask: are adjacent cells alike?

If we’re exercising statistical integrity, we actually count and measure and calculate, to determine the likelihood of whether the pattern could happen randomly.

I’m sure you’ve heard the phrase, “Hypothesis testing.” When we do that, we’re testing the hypothesis that something, like this pattern, happened by accident.

For many of our maps, the patterns are so obvious that we do the hypothesis testing in our heads. But for this ambiguous spatial pattern, we can measure adjacency: how many black squares are near other black squares? More than we would expect would happen randomly?

So, how do we ask that?



Measuring trends across space…

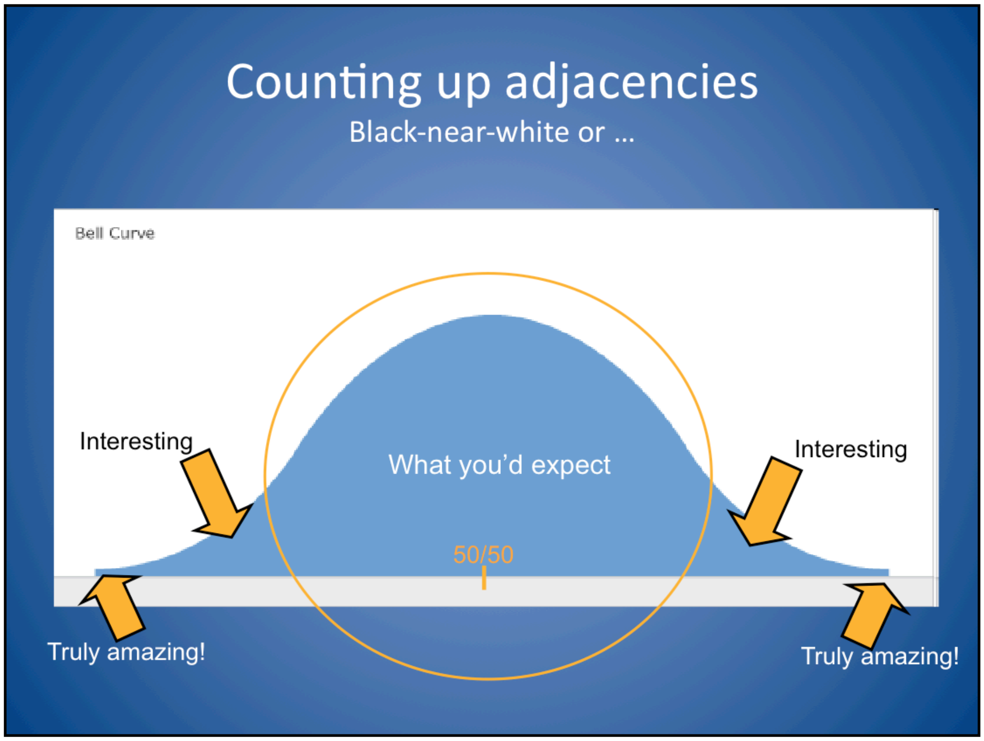
Are adjacent cells alike?

- Is there an even, balanced proportion of:
  - – Black near Black?
  - – White near white?
  - – Black near white?
  - – White near Black?

Figure is half & half, black & white

If we count up the adjacencies in this grid, these are pretty evenly distributed. So I'm going to say that there is a high probability that this happened randomly.

Now, before you get all hot under the collar because you don't agree with me, I'm not saying this pattern DIDN'T happen for a reason. Maybe it did. Maybe there were *several* confounding reasons influencing this distribution. We don't know. We can't know just by looking at this. The only thing we can say is that there's not enough clustering going on for us to say that something caused this black area down here or the white area over here.



Counting up adjacencies

Black-near-white or …

If you've taken statistics, then you'll recognize the bell curve. It's the iconic representation for "most things that happen are pretty normal." It's also called the normal curve.

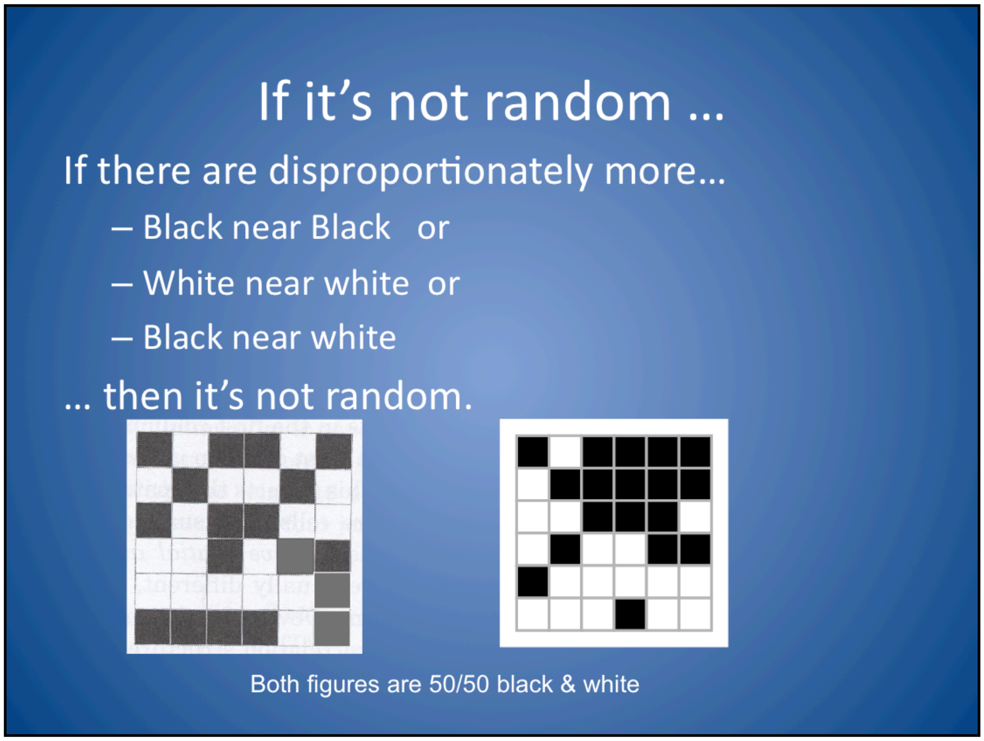
When we pulled those 36 tiles out of the bowl, we're expected to get pretty close to a 50/50 split, whether a tile is next to it's own color or next to the opposite color.

It actually could happen for you to have one tenth of the black tiles surrounded by white tiles, but that would be *very* interesting. If you ended up with the even split black/white pattern that we saw in the beginning, that would be a case for the Guiness Book of Records. Same thing for making a perfect checkerboard.

It could conceivably happen, but it's such a remote chance that we say there is ZERO probability that it ever would.

And we can say that we're pretty darn sure that these "interesting" cases wouldn't happen by accident. It's all about a continuum of certainty.

If we want to be sure something is a meaningful cluster, we want it to be in the "interesting" area … not a pattern that could very likely happen for no particular reason.



If it’s not random …

If there are disproportionately more…

- Black near Black   or
- White near white or
- Black near white

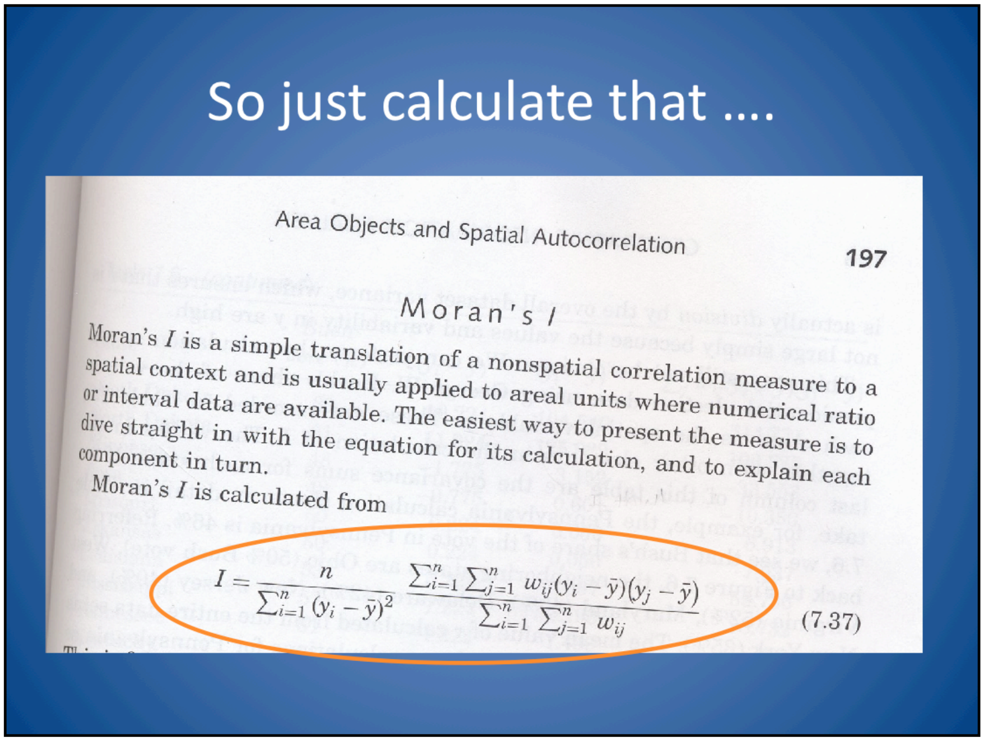
… then it’s not random.

Both figures are 50/50 black & white

So here are two figures that are each half and half, black and white.

Here’s our previous pattern on the left and a new pattern on the right.

It’s my personal opinion that we have a meaningful cluster up here. But we can just measure it and do the calculations….



So just calculate that ….

Area Objects and Spatial Autocorrelation 197

Moran's I

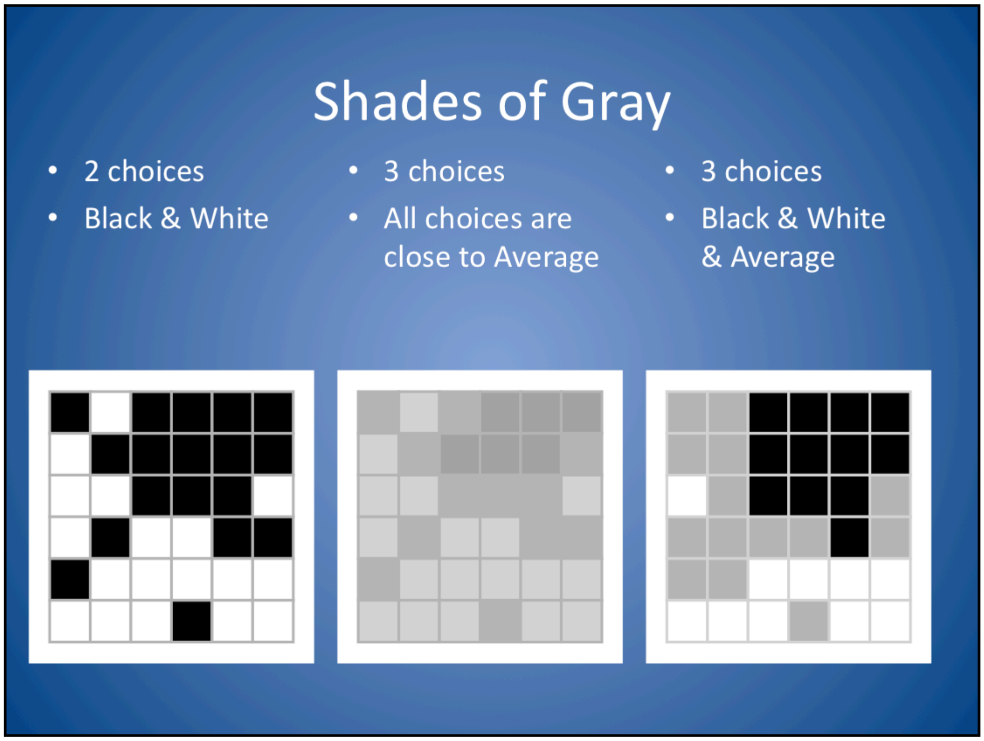
Moran's I is a simple translation of a nonspatial correlation measure to a spatial context and is usually applied to areal units where numerical ratio or interval data are available. The easiest way to present the measure is to dive straight in with the equation for its calculation, and to explain each component in turn.

Moran's I is calculated from

$$I = \frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \qquad (7.37)$$

Hey, no problem, right? We'll just use this little calculation to figure it out.

Or, as a refreshing alternative, we can let the software do it. What do you think?



Shades of Gray

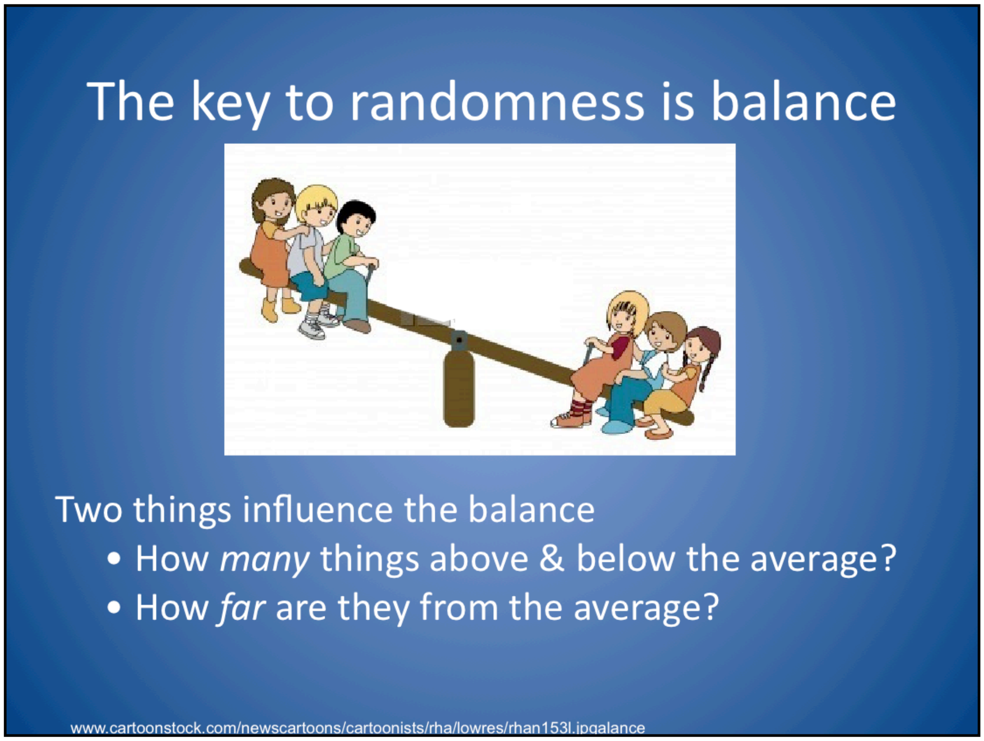
- 2 choices
- Black & White

- 3 choices
- All choices are close to Average

- 3 choices
- Black & White & Average

But before we move to maps and the software tools, let’s expand a bit, from black and white to shades of gray.

….



The key to randomness is balance

Two things influence the balance

- How *many* things above & below the average?
- How *far* are they from the average?

www.cartoonstock.com/newscartoons/cartoonists/rha/lowres/rhan153l.jpgalance

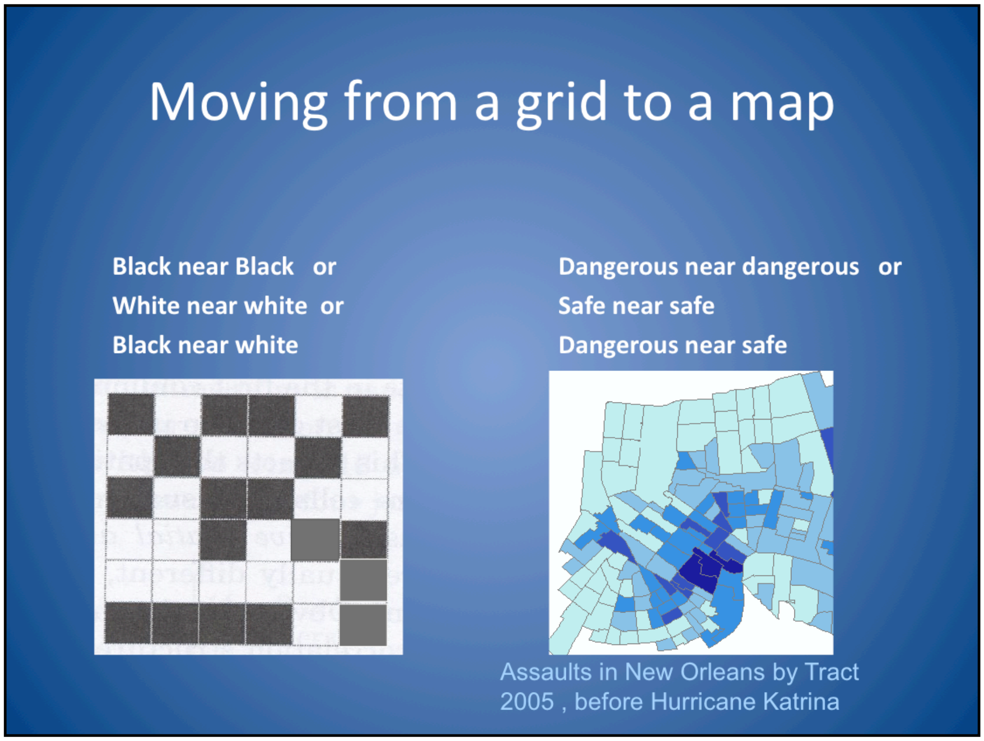
With the assault data, we don’t have a simple count … we want to find a balance.

First we need to find the balance point … the Mean, aka the average.

We count the number “above average” tracts near other “above average” tracts.
And we count the number of “*below* average” near other “*below* average” tracts.

And then we take into account in how FAR above (or below) average they are.
That is, if we have 2 tracts that are both way above average, that counts more than 2 tracts that are just barely above average.

And if we have one really dangerous tract next to a really safe tract, then that’s more meaningful than two tracts that are indeed above and below the average, but only by a mugging or two.



Moving from a grid to a map

Black near Black or
White near white or
Black near white

Dangerous near dangerous or
Safe near safe
Dangerous near safe

Assaults in New Orleans by Tract
2005 , before Hurricane Katrina

Actually, before we get to the software, I want to talk about the case where we look at more than just black and white / yes and no. When my student wanted to look at hot spots, we were looking at “tracts with lots of assaults” and “tracts with few assaults”. I dare say every census tract in New Orleans had at least one, those poor people.

So let’s move from the familiar image on the left to the map on the right.

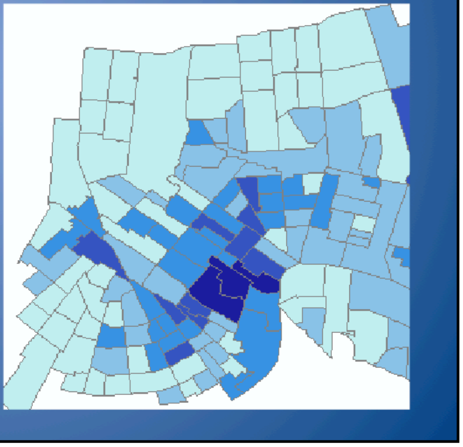
The dark blue areas had more assaults. I’m calling them “more dangerous”, although I didn’t normalize by area or by population – this is just a map of the simple count of assaults in 2005, prior to Hurricane Katrina.

It looks like there is a dark patch in the middle there. Is it really a cluster? Or could that pattern have happened randomly?

Is It Random?

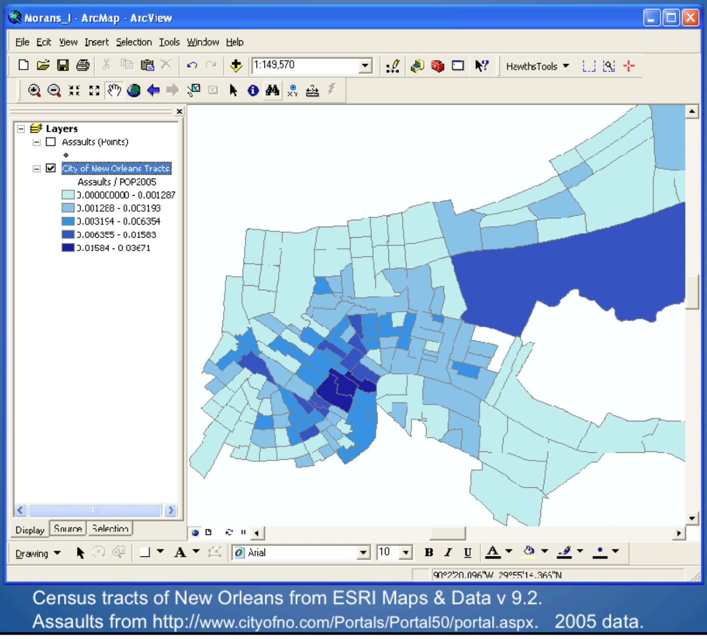
If there are disproportionately more…

- Dangerous near dangerous
- Safe near safe
- Dangerous near safe



Then it's not random

Random or Meaningful Clusters?


Morans_I - ArcMap - ArcView
File Edit View Insert Selection Tools Window Help
1:149,570
HawthsTools
Display Source Selection
Drawing
Arial
10
B I U

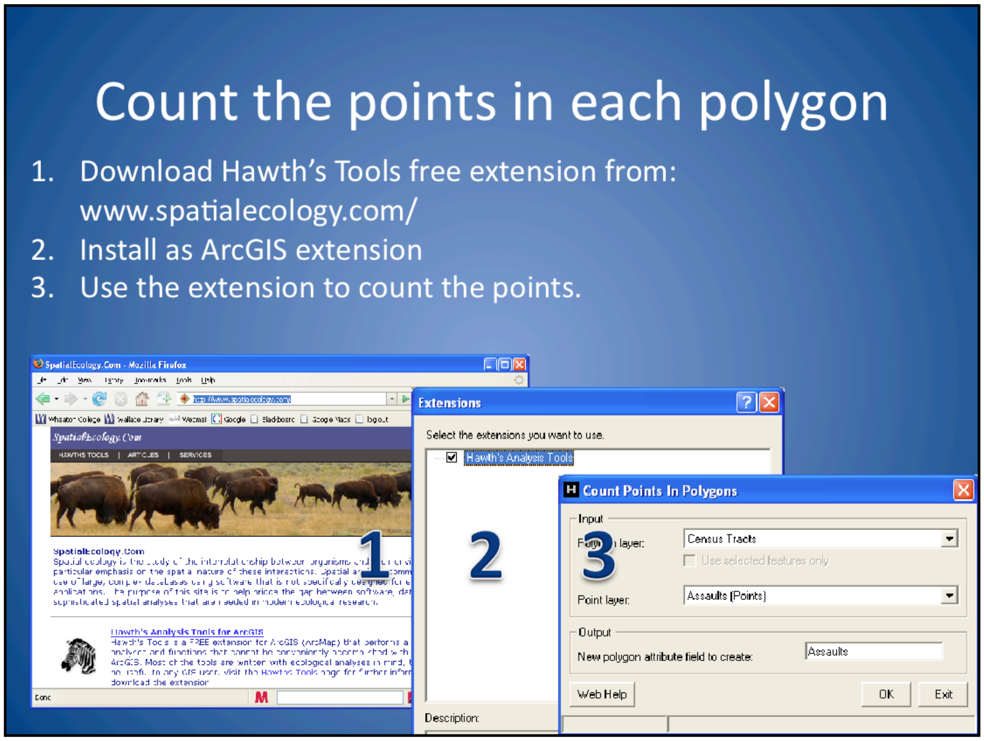
Layers
- ☐ Assaults (Points)
- ☑ City of New Orleans Tracts

Assaults / POP2005

🟦	0.00000000 - 0.001287
🟦	0.001288 - 0.003193
🟦	0.003194 - 0.006354
🟦	0.006355 - 0.01583
🟦	0.01584 - 0.03671

Census tracts of New Orleans from ESRI Maps & Data v 9.2.
Assaults from http://www.cityofno.com/Portals/Portal50/portal.aspx. 2005 data.

We did this with a free extension to ArcGIS called Hawth's tools, from SpatialEcology.com



Count the points in each polygon

1. Download Hawth's Tools free extension from: www.spatialecology.com/
2. Install as ArcGIS extension
3. Use the extension to count the points.

Extensions

Select the extensions you want to use.

☑ Hawth's Analysis Tools

Count Points In Polygons

Input

Polygon layer: Census Tracts

□ Use selected features only

Point layer: Assaults (Points)

Output

New polygon attribute field to create: Assaults

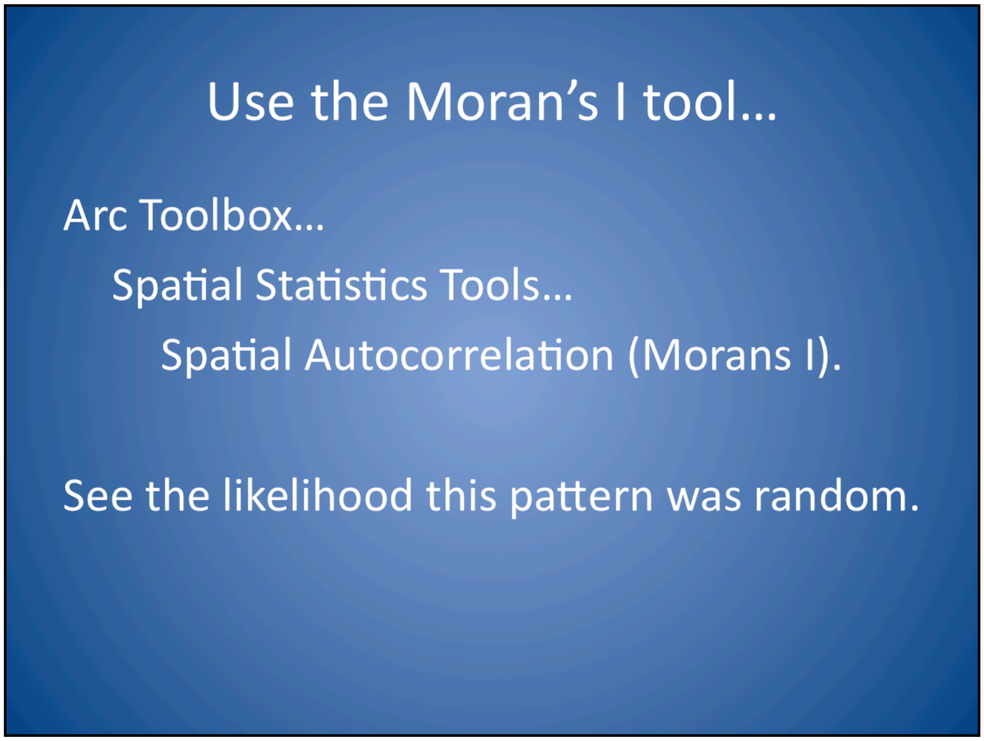
SpatialEcology.Com - Mozilla Firefox
SpatialEcology.Com
HAWTHS TOOLS | ARTICLES | SERVICES
Hawth's Analysis Tools for ArcGIS
1
2
3
Description:
Web Help
OK
Exit

So now we're getting to the software part. This is all in the handout.

Step 1 is to download the free extension

Step 2 is to install Hawth's tools in ArcGIS

Step 3 is to use the dialogue box to specify the base map and the points you want counted. Here I'm asking it to start with the census tract shape file, then count how many points are in each census tract. It adds an additional column to the end of the data table, with that count.



Use the Moran’s I tool…

Arc Toolbox…

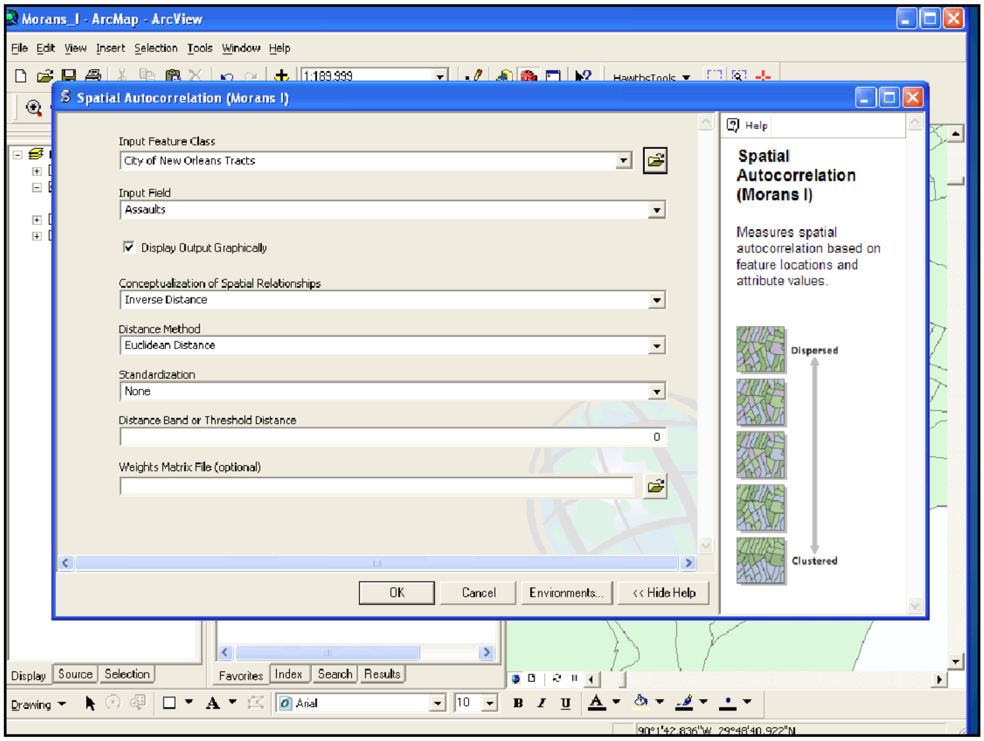
Spatial Statistics Tools…

Spatial Autocorrelation (Morans I).

See the likelihood this pattern was random.

Now we have the map equivalent of our little grid of black and white squares.

So we go into the ArcToolbox, choose Spatial Statistics Tools, then Spatial Autocorrelation, and let it crank.


Morans_I - ArcMap - ArcView
File Edit View Insert Selection Tools Window Help
1:189,999
HawthsTools
Display Source Selection
Favorites Index Search Results
Drawing
Arial
10
B I U
90°1'42.836"W 29°48'40.922"N
OK
Cancel
Environments...
<< Hide Help
Help
Dispersed
Clustered

Spatial Autocorrelation (Morans I)

Input Feature Class: City of New Orleans Tracts

Input Field: Assaults

☑ Display Output Graphically

Conceptualization of Spatial Relationships: Inverse Distance

Distance Method: Euclidean Distance

Standardization: None

Distance Band or Threshold Distance: 0

Weights Matrix File (optional)

Spatial Autocorrelation (Morans I)

Measures spatial autocorrelation based on feature locations and attribute values.

Here is the dialogue box for the Spatial Autocorrelation tool.
This is in the handout, too.

We tell it which shape file we want to look at, and the name of the new column we just made, that counts how many assaults in each tract.

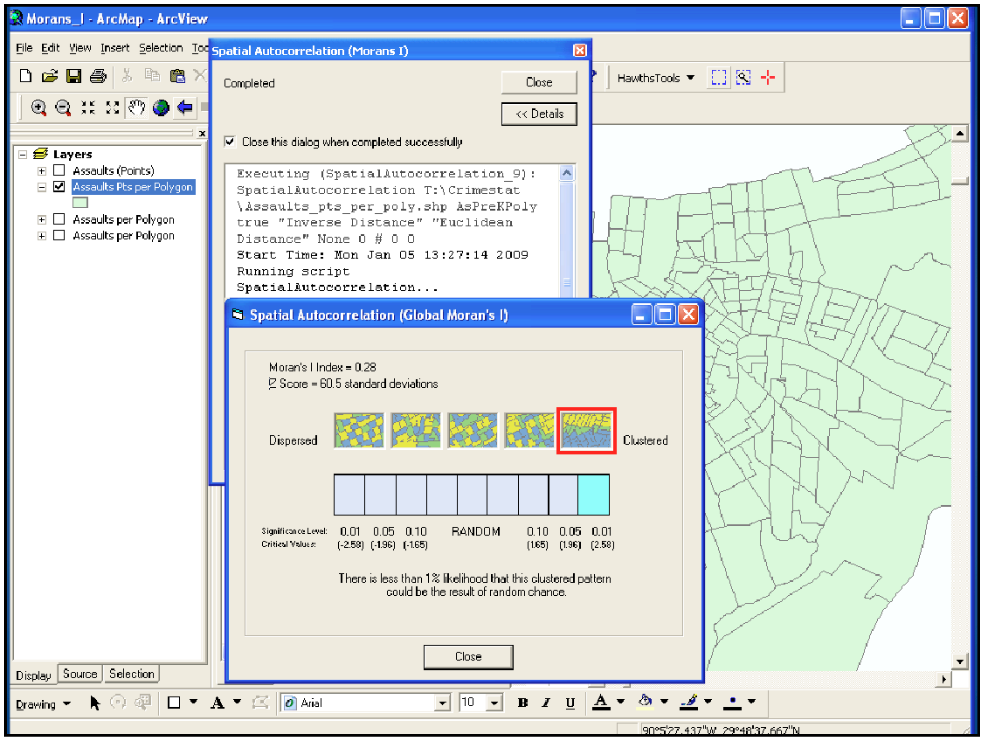
I took all the defaults because I don't know any better (and nothing else worked). I only know about a nickel's worth about any of these options, but if you're curious I'm happy to share that.

On the right ESRI nicely gives us the range of outcomes we might get. On the bottom we see an example of strong clustering. Blues next to blues, greens next to greens. This is analogous to our grid with all the whites on one side and all the blacks on the other.

On the top we see non-clusters. There are NO blues next to blues or greens next to greens. This is like our checkerboard. The example I imagined where we might see that in real life would be cardinal's nests. Cardinals are very territorial, so they would never choose to put their nests in neighboring yards. They're going to put them as far apart from each other as they can get. This is non-random, like the checkerboard pattern.

And, of course, it's all a continuum of probability.

I



Spatial Autocorrelation (Morans I)

Completed

Close

<< Details

☑ Close this dialog when completed successfully

```
Executing (SpatialAutocorrelation_9):
SpatialAutocorrelation T:\Crimestat
\Assaults_pts_per_poly.shp AsPreKPoly
true "Inverse Distance" "Euclidean
Distance" None 0 # 0 0
Start Time: Mon Jan 05 13:27:14 2009
Running script
SpatialAutocorrelation...
```

Spatial Autocorrelation (Global Moran's I)

Moran's I Index = 0.28
Z Score = 60.5 standard deviations

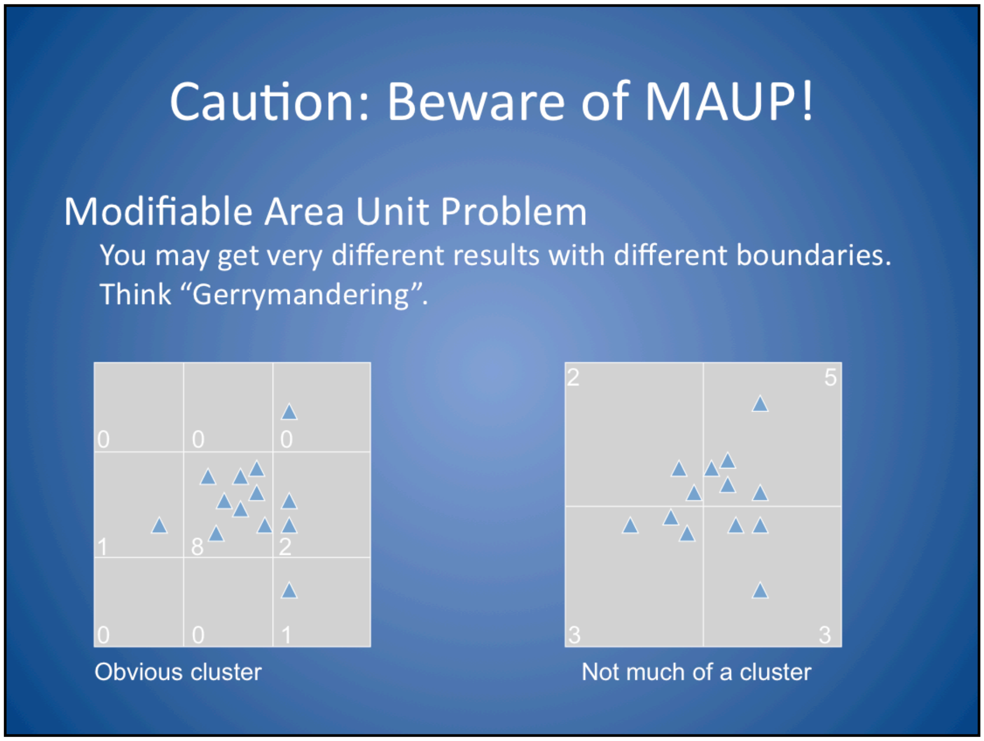
Significance Level:	0.01	0.05	0.10	RANDOM	0.10	0.05	0.01
Critical Values:	(-2.58)	(-1.96)	(-1.65)		(1.65)	(1.96)	(2.58)

There is less than 1% likelihood that this clustered pattern could be the result of random chance.

Close

And then the tool cranks away and pops up a little display and …
Ladies and gentlemen, we have a winner! Our data is strongly clustered. It’s extremely unlikely that we’d see that pattern by accident.

We see that the “cardinal nest” example of non-random dispersion over here, and the “assault” example of non-random dispersion over here, and between them are all the random things that happen for no particular reason.



Caution: Beware of MAUP!

Modifiable Area Unit Problem

You may get very different results with different boundaries. Think "Gerrymandering".

Obvious cluster

Not much of a cluster

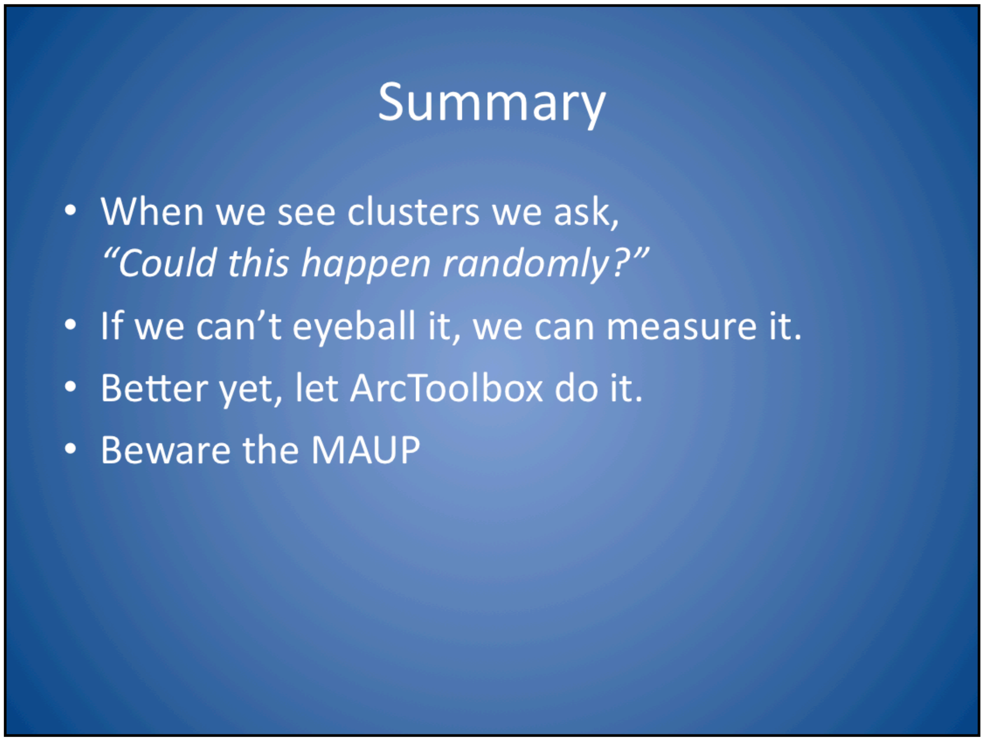
…

We always need to be aware of the Modifiable Area Unit Problem when we convert point data to polygon data. Take a look at this spatial distribution. We see a very clear cluster of points.

But look what can happen if we start counting the points inside a polygon, to turn it from points to polygons.

If we decide to divide our area into 9 polygons, then our cluster is preserved .. .it shows off very nicely as an 8 among much smaller numbers.

If we use 4 polygons, though, we effectively obscure the cluster by sharing it among all the polygons. 5 just isn't that much bigger than 3.



Summary

- When we see clusters we ask, *"Could this happen randomly?"*
- If we can't eyeball it, we can measure it.
- Better yet, let ArcToolbox do it.
- Beware the MAUP

So, to summarize. ….

Clusters: Random or Meaningful?

Using Moran’s I Test to measure

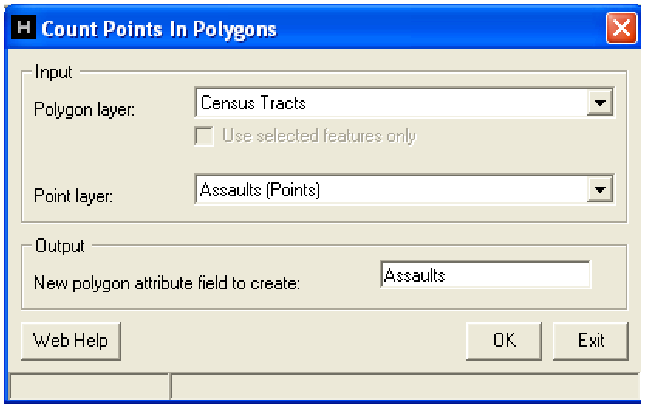
1.To turn point data into polygon data, begin by downloading
Hawth’s Analysis Tools extension from www.spatialecology.com

2. Install the extension into ArcMap

3. Make the extension toolbar visible

4. Choose Hawths Tools /Analysis/Count Points in Polygons

5. Fill in dialog box.
In this example, we count the number of assault points in each census tract.
We add a new field to the Census Tract attribute table called Assaults.
The new field contains the number or assaults in each census tract.



Count Points In Polygons

Input

Polygon layer: Census Tracts

☐ Use selected features only

Point layer: Assaults (Points)

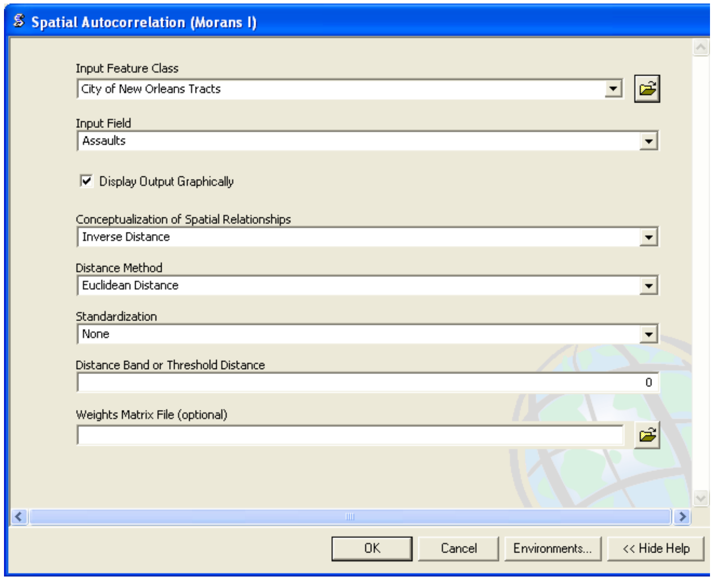
Output

New polygon attribute field to create: Assaults

6. Export the layer to make a fresh file.

7. Open ArcToolbox / Spatial Statistics Tool / Spatial Autocorrelation (Morans I)

8. Fill in the dialog box.
In this example, we measure the degree of randomness in the field
called Assaults, in the attribute table of City of New Orleans Tracts.
Check the “Display Output Graphically” box.
Take the defaults for the other options.



Spatial Autocorrelation (Morans I)

Input Feature Class: City of New Orleans Tracts

Input Field: Assaults

☑ Display Output Graphically

Conceptualization of Spatial Relationships: Inverse Distance

Distance Method: Euclidean Distance

Standardization: None

Distance Band or Threshold Distance: 0

Weights Matrix File (optional)

9. The resulting output tells the probability that the pattern is random.
On the left is non-random “dispersed” like cardinals’ nests and checkerboards.
On the right is clustered, like crimes and boutiques.

J.Lund & A.Anderson

January 9, 2009