Deck 11 - Geostatistics

Raster and Boolean operations recap, geostatistics and spatial statistics

---

Intro to GIS – UMass Amherst – Michael F. Nelson

Start Zoom Recording!

Reminder for Mike

Overview

Boolean Algebra and Rasters Recap

- Multiplication vs. AND
- Addition vs. OR
- Raster Calculator and Reclassify Output

Geostatistics and Spatial Statistics

- Probability Distributions
- Point Patterns
- Spatial Autocorrelation
- Interpolation

Final Projects

- Questions

Important End-Of-Course Info

Ignore at your own peril!

Finishing up the course assignments

- Remember that your lowest lab grade is dropped.
- Recall that there are no penalties for late assignments, but late submissions may have delayed grading.
- All lab assignments must be submitted by **midnight July 14th**
  - This due date is not flexible except under extenuating circumstances arranged beforehand.
  - Be in touch ASAP if you have concerns.
  - If you wait until the last minute, we are unlikely to be able to help.
  - Note that

Final Projects: Methodological Outline

- We're still missing a lot of submissions.
- If you haven't completed this assignment, you will need to do so ASAP. It will be very difficult for you to produce an effective poster without feedback on this assignment.
- The hard deadline is **midnight July 14th**
- If you're struggling, reach out to myself or the TA immediately. If you rush through this, you won't get a good grade and you will not have a firm basis for a successful poster.
- Also… We can't accept methodological outlines submitted after the final poster. You need to do these steps in order!

Midterm Earn-Back

- This assignment uses the same datasets as the midterm.
- Hard deadline is last day of classes (July 14th)
- You can make up over 50% of the midterm points (but not earn over 100%).

Final Poster

- Final poster preferred deadline is Monday July 10th.
  - Submissions received by midnight on the 10th will receive feedback and an opportunity for revision and resubmission.
- Final posters are due by the end next week: July 14th.
  - This date is not negotiable, unless under extenuating circumstances that have been arranged in advance.
- Submit your final poster as a .pdf on Moodle.
- Make sure your maps are exported in high resolution.

Final Poster: Optional Virtual Poster Session

- Final poster session is not required; however, you'll earn 2% extra credit applied to your final course grade if you present your poster at the virtual poster session.
- If you want to participate in the virtual session, indicate your availability on the When2Meet poll in Moodle.

Resources

- Please continue to email both me and your TA with questions.
  - To get the quickest response, write to both of us!
- My availability for outside-of-class meetings is sometimes limited, especially on weekends.
  - Mea culpa, but sometimes I miss an email. Please don't hesitate to send a follow-up reminder if I don't reply right away.
- Your TA is an incredible resource – use their knowledge!
  - You probably already know this, but your TA's contact info is in Moodle.
  - TA office hours are Thursday 11 – 1.

Due Date Summary – Hard Deadlines

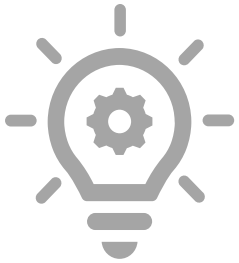
- Hard deadline for all assignments is Friday, July 14th.
- Deadline to receive feedback and an opportunity for revision/resubmission on your final project is Monday, July 10th.

Model Thinking

Before we delve into the weirdness that is statistics – don't forget about model thinking

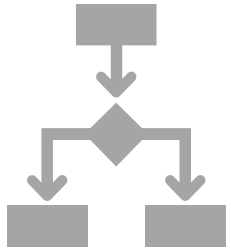
A model is a simplified abstraction of reality.

"Everything **should** be made as **simple as possible**, but no simpler."



All models make assumptions (known or unknown).

When we propose a model, we've assumed that it is a useful way to think about reality!



We want to create models that are helpful for our purpose.

The goal of our model is to help us understand a particular set of questions.

A map is a cartographic model, what components o you control to make your map helpful?

Statistical models have two parts: a deterministic model and a stochastic model.

Deterministic functions produce the same output with the same input.

Stochastic processes can produce different outputs from the same inputs

The Dual-Model Paradigm

Deterministic Model: like a mathematical function, it always produces the same output with the same input. *The model of average or expected behavior.*

Stochastic Model: this models the range of possible variations that we could expect to see. *The model of variability or noise.*

Stochastic Process: a procedure whose outcome is uncertain. *Given the same inputs, a range of outcomes is possible.*

[Frequentist] Statistics

Philosophy, Assumptions, Hypothesis Testing

Bird's eye view of the assumptions and philosophy of Frequentist Statistics

Sample

- A sample is a **finite subset** of the population.
- We hope to estimate the unknowable population parameters via our **sample statistics.**

Population

- A population is **infinitely large**, and its properties are **unknowable.**
- A population has real properties: **population parameters** (mean, variance, etc.) but we can never know their exact values.

Bird's eye view of the assumptions and philosophy of Frequentist Statistics

Repeated Sampling

- Frequentism relies on the concept of hypothetical (infinite) repeated sampling.
- Sampling is a stochastic process.
- We aim to estimate **population parameters** using **sample statistics.**

Uncertainty and Frequentist Weirdness

Significance/confidence based on repeated sampling:

- Not correct: "I'm 95% sure the population value is between 3 and 5."
- Correct: "If I repeated my sampling process many times, the true population value would fall within my interval 95% of the time"

Null Hypotheses

A null distribution can tell us what we would expect to observe if data were generated **randomly**.

Null hypothesis: different perspectives

There is no relationship between X and Y	Knowing the value of X **tells us nothing** about the value of Y	X and Y are **independent**

Alternative Hypotheses

What we think is **actually happening**.

This is described by an **alternative distribution.**

Alternative hypothesis: different perspectives

There is a positive relationship between X and Y	If we know the value of X is high, we can guess that Y should be also be high.	If we know X, we can make a better guess about Y	X and Y **non-independent.**	X and Y vary in a coordinated manner.

Null hypotheses and the dreaded/abused **p-value**

- The p-value is ever-present in Frequentist statistics, but what is it?

Author unknown…. or unknowable?

P-value: Intuitive perspectives

The p-value is the probability that the pattern we observed in our data ***occurred by chance alone***.

It is the probability that we observed something interesting if the ***null hypothesis were true***.

It is the ***false-positive*** rate.

Statistical Inference

- Trying to make an educated guess about a population from a sample.
- Many different inference techniques and paradigms
- Different sets of assumptions, and ways to violate assumptions
- Lots of theoretical and software tools

Frequentism isn't the only way!

Bayesian statistics

Machine learning techniques

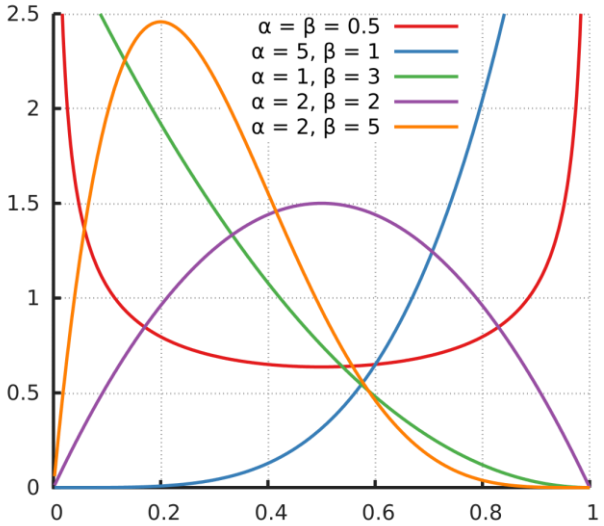
Simulation

Maximum Likelihood methods*

*ML methods are also used in other paradigms

Probability Distributions

What is a probability distribution?



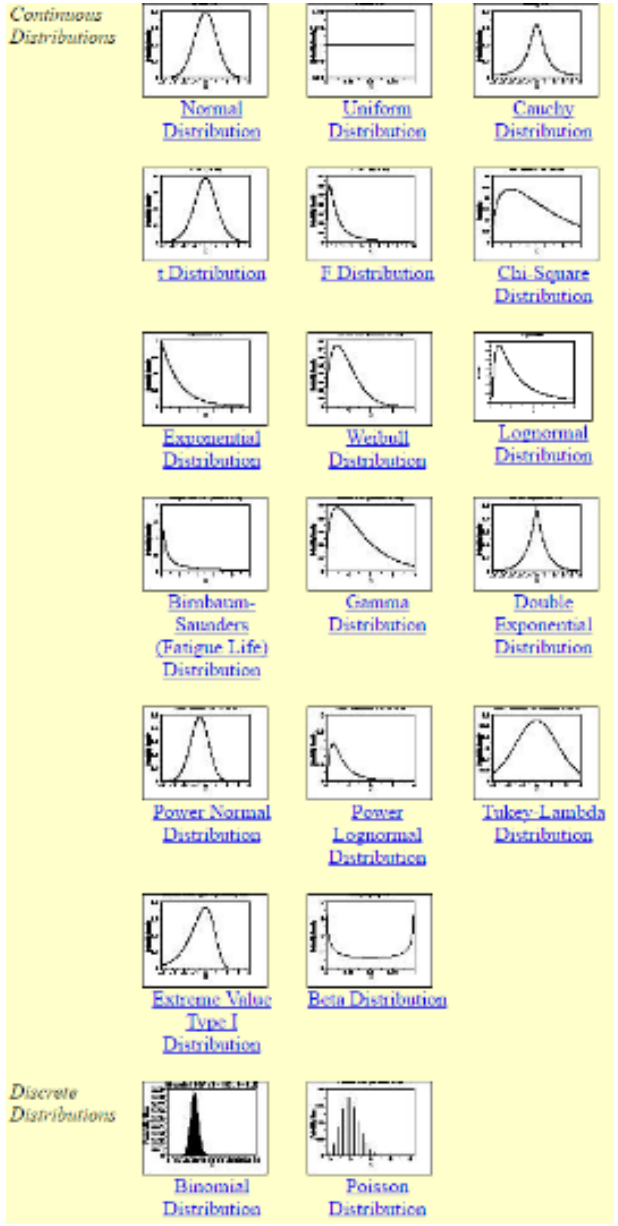
Beta distributions

Thank you, Microsoft Bing!

- A theoretical probability distribution maps an event to the relative likelihood, compared to all other possible events.
  - What is the probability of rolling a 6 on a fair die?
  - How many ways are there to observe a sum of 12 when you roll 2 fair dice?
  - What is the probability of getting a sum of 7 when you roll 2 fair dice?
- Which distributions do you already know about?

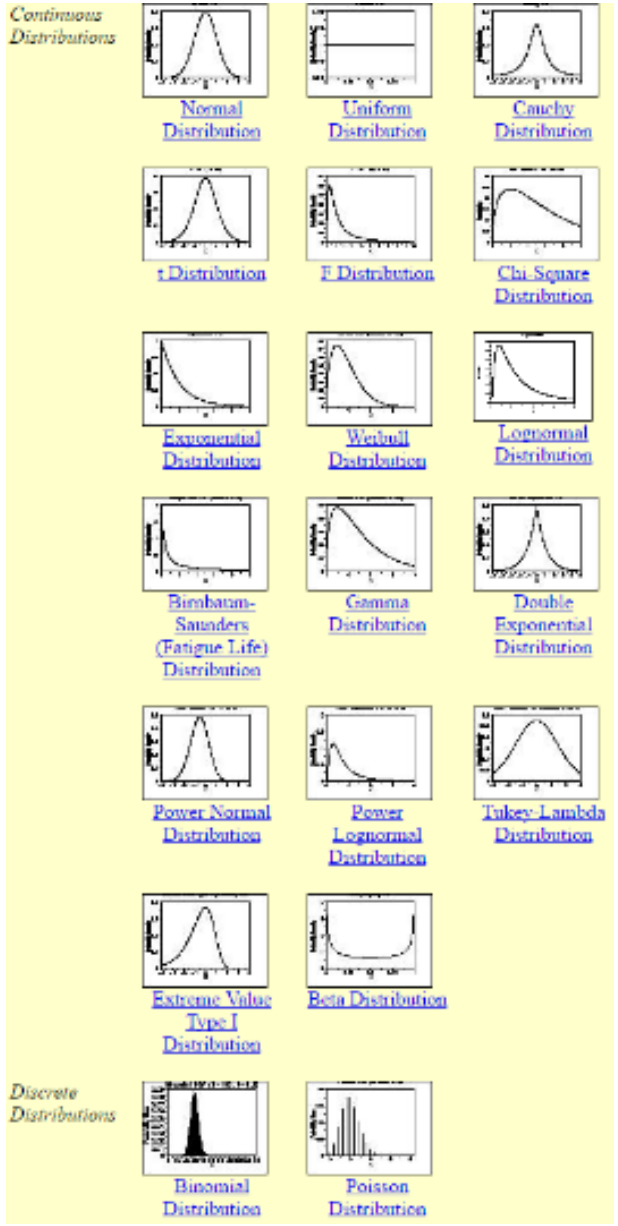
Distributions

- Help us understand the arrangement and properties of data.
- Can help us make predictions.
- Can help us determine whether an event is unlikely.
- Are defined by one or more **parameters.**
  - Normal dist. Parameters are mean and standard deviation



Types of Distribution

- There are hundreds of named distributions!
- Most of us have heard of the Normal distribution.
- It describes a **continuous quantity**
  - It is symmetrical
  - It occurs frequently in theory and practice
- **Discrete distributions**
  - Models for counts, can only take on **integer values.**
- Some distributions help us describe spatial things!

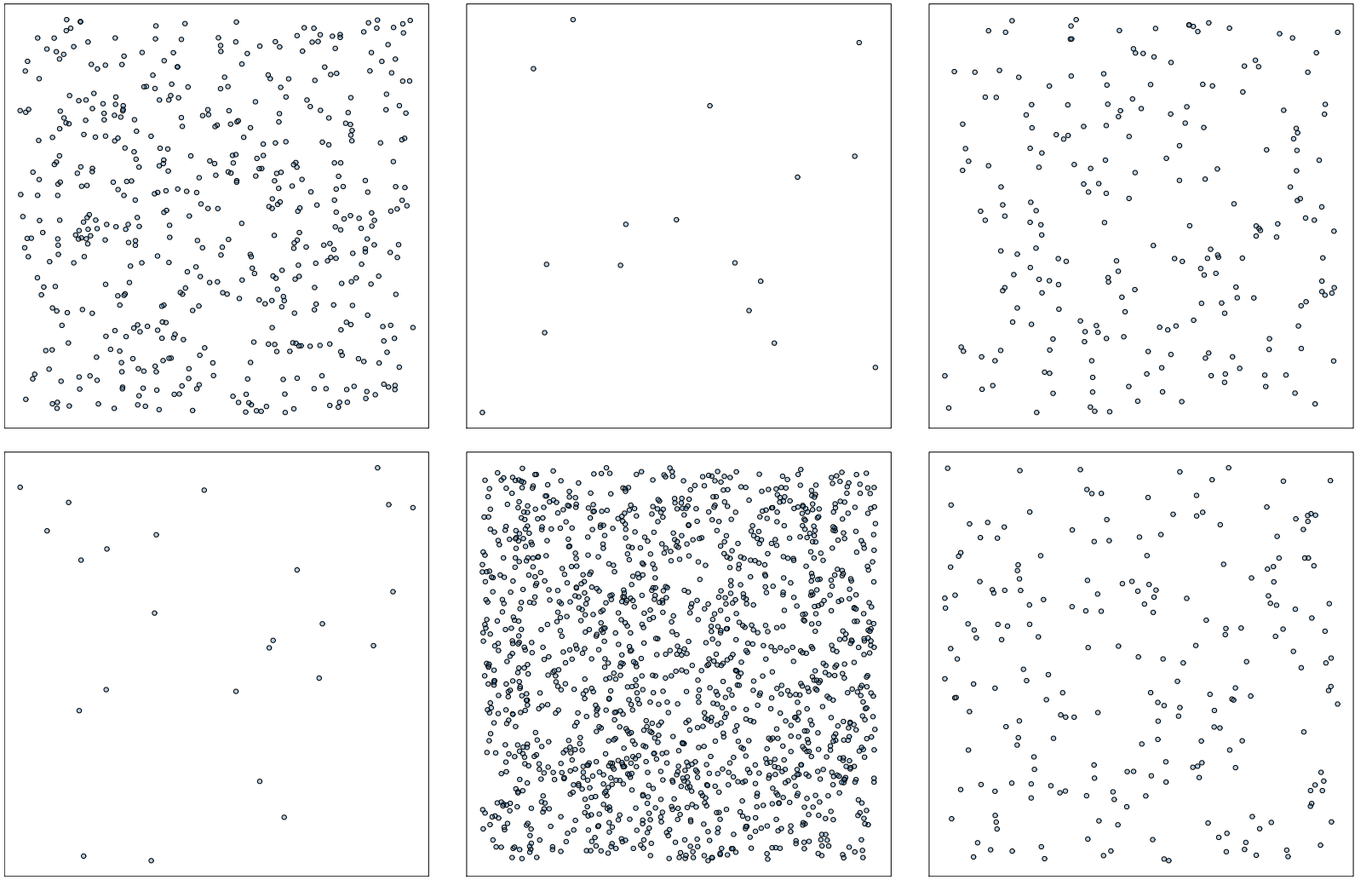


Spatial Point Patterns

Point Distributions

- What would we expect a completely random point pattern to look like?
  - Evenly spaced?
  - Irregular?
  - Clustered?
- Is there a distribution that describes **Complete Spatial Randomness**?

Which of these patterns are random?

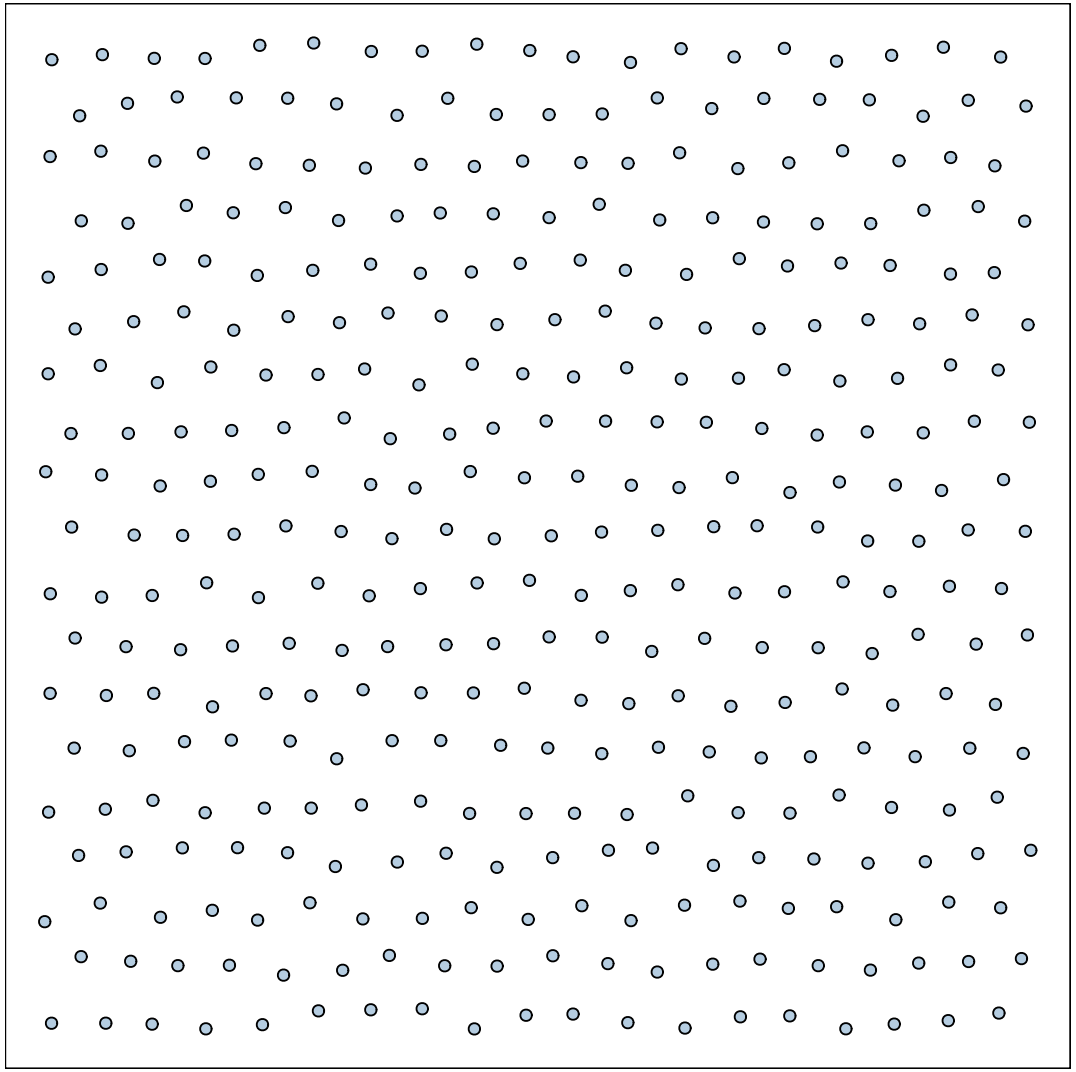


Regular Spacing: Overdispersion

Regularly spaced patterns can result from:

- Repulsive processes

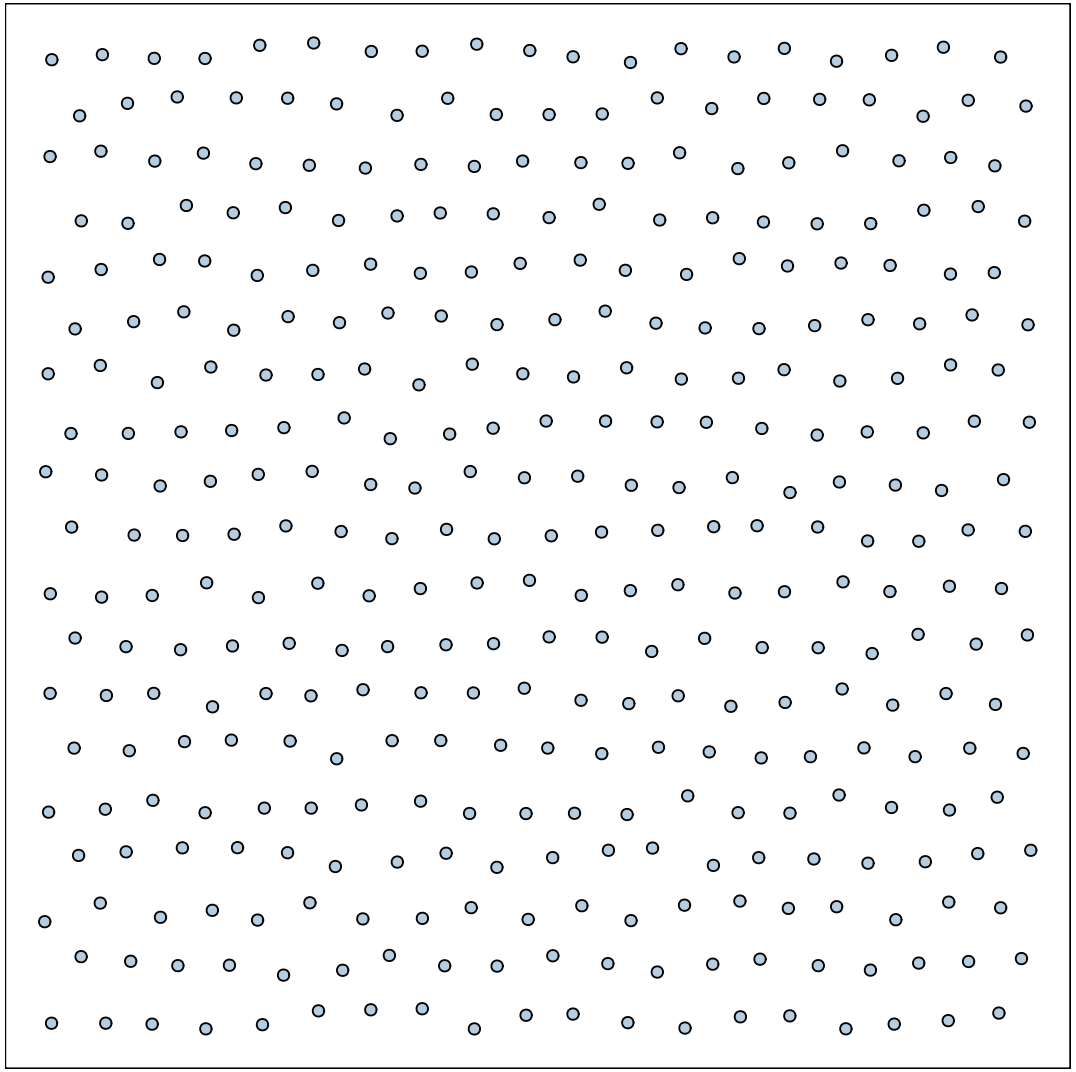
- Anthropogenic influence



Regular Spacing: Overdispersion

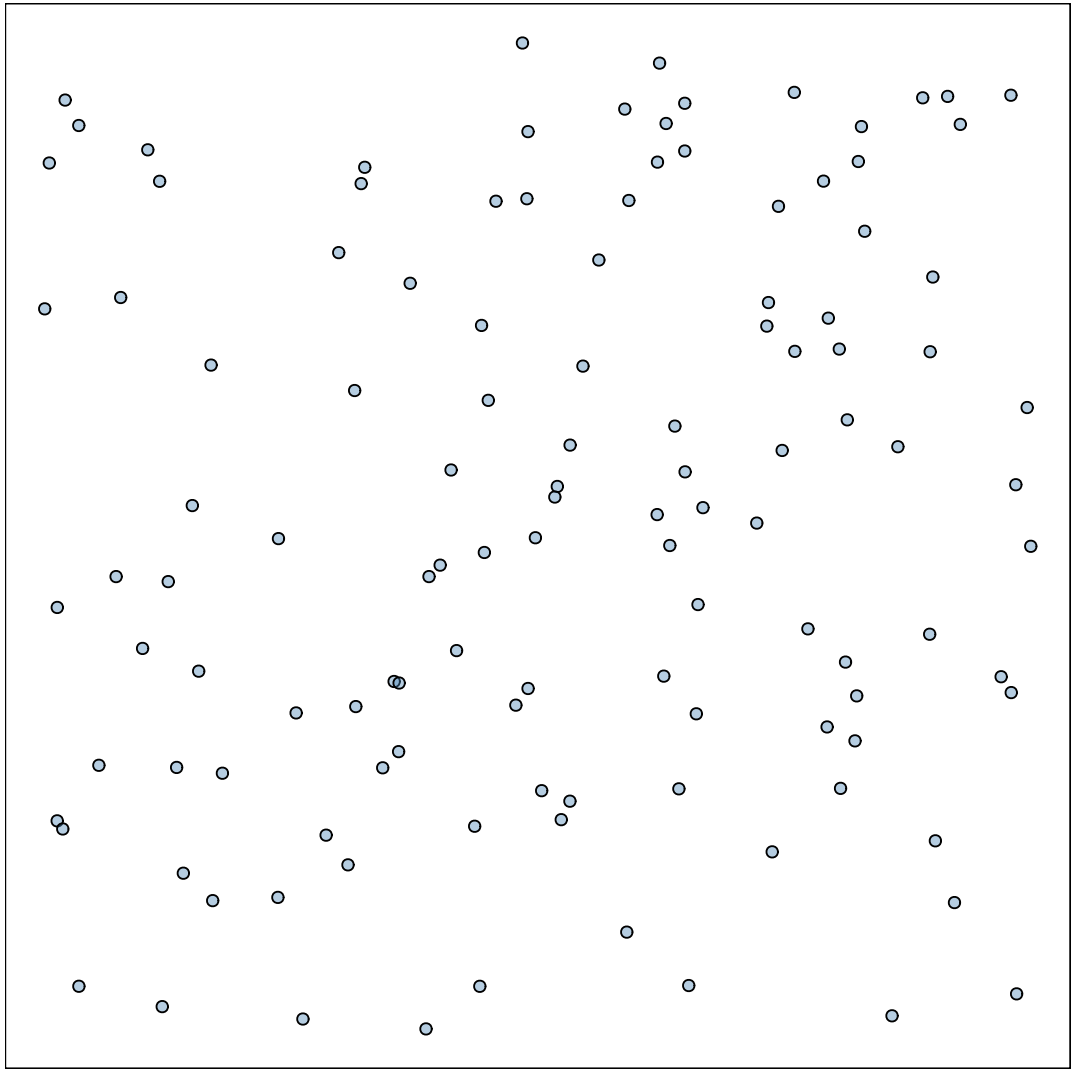
Regularly spaced patterns can result from:

- Repulsive processes
  - Inhibition
  - Competition for resources
- Anthropogenic influence
  - Streets on a grid
  - Houses in a neighborhood
  - Farmers' fields

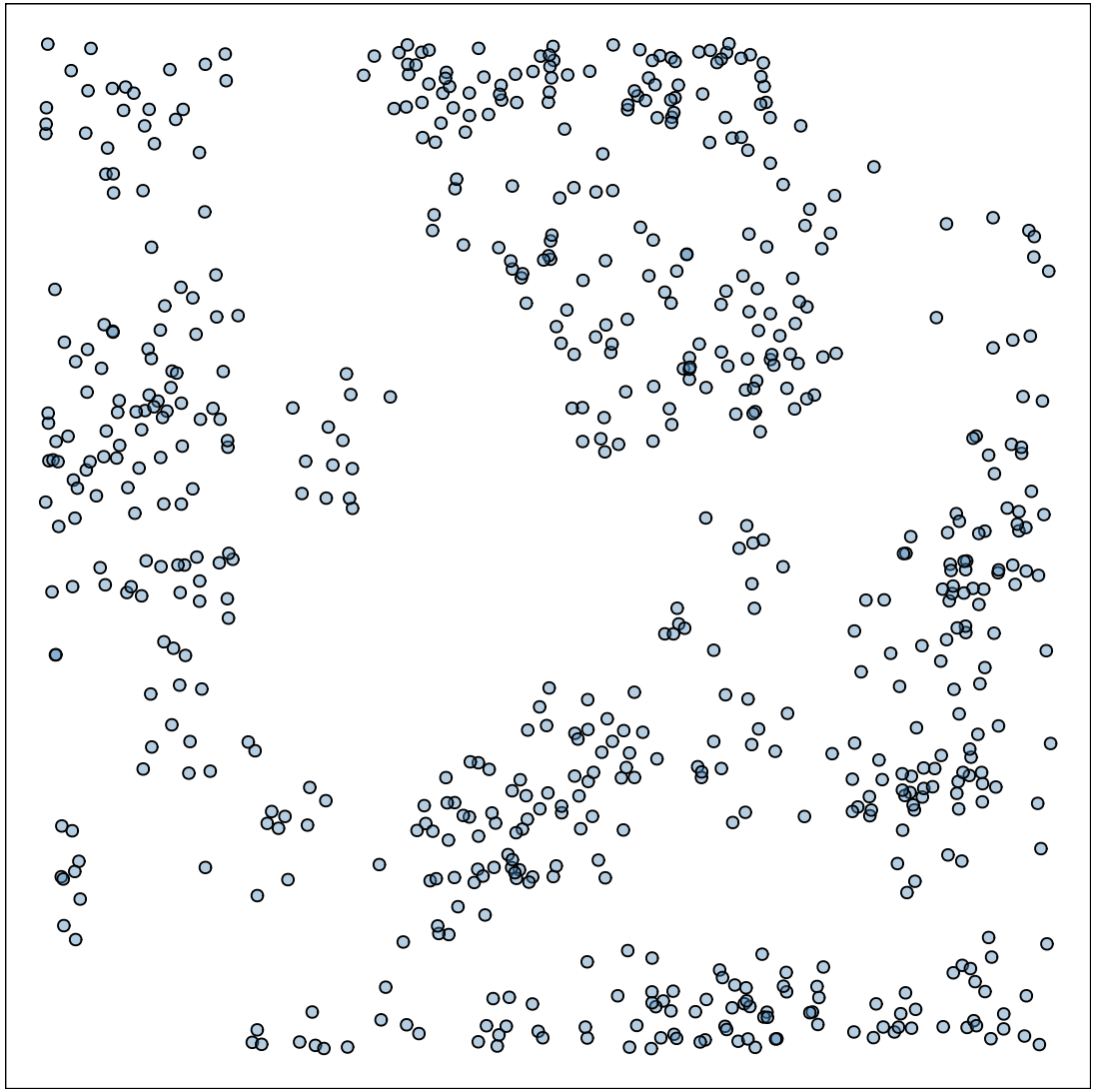


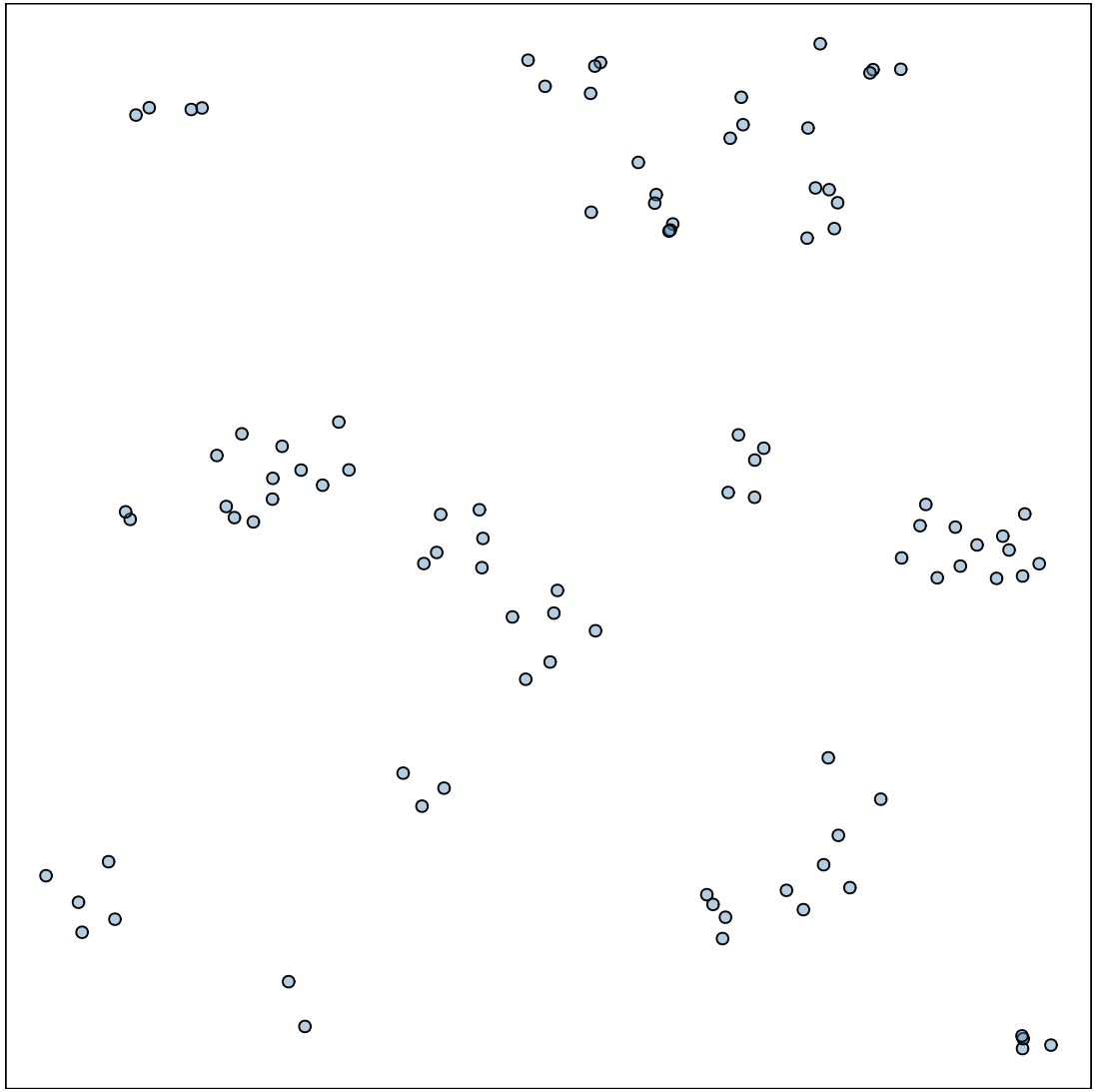
Complete Spatial Randomness (CSR)

- Randomly spaced points result from processes in which events are independent.
- CSR can look less ‘random’ than we expect.
  - Remember, we’re [too] good at seeing patterns.
- CSR is described by the Poisson distribution



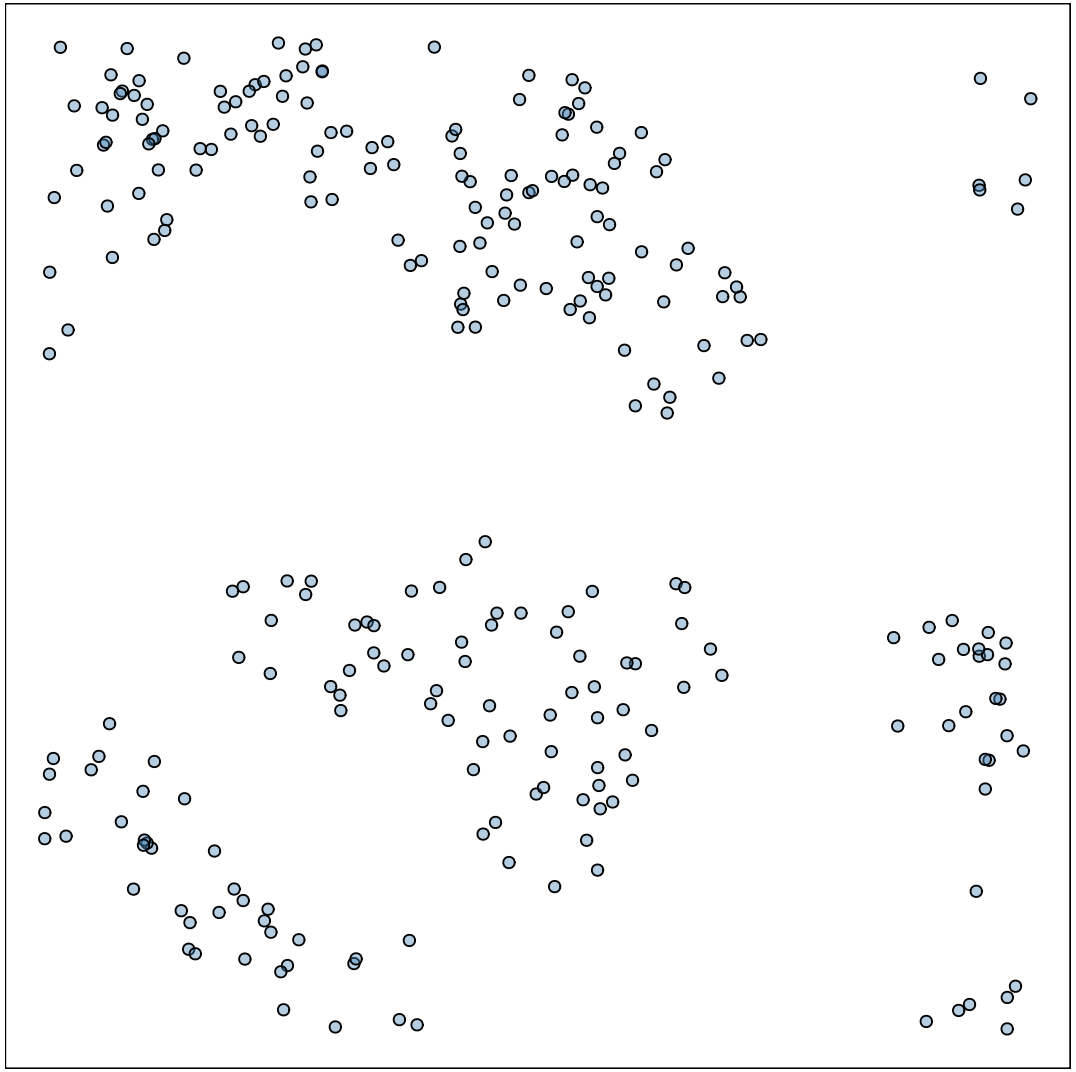
Clustered Point Patterns





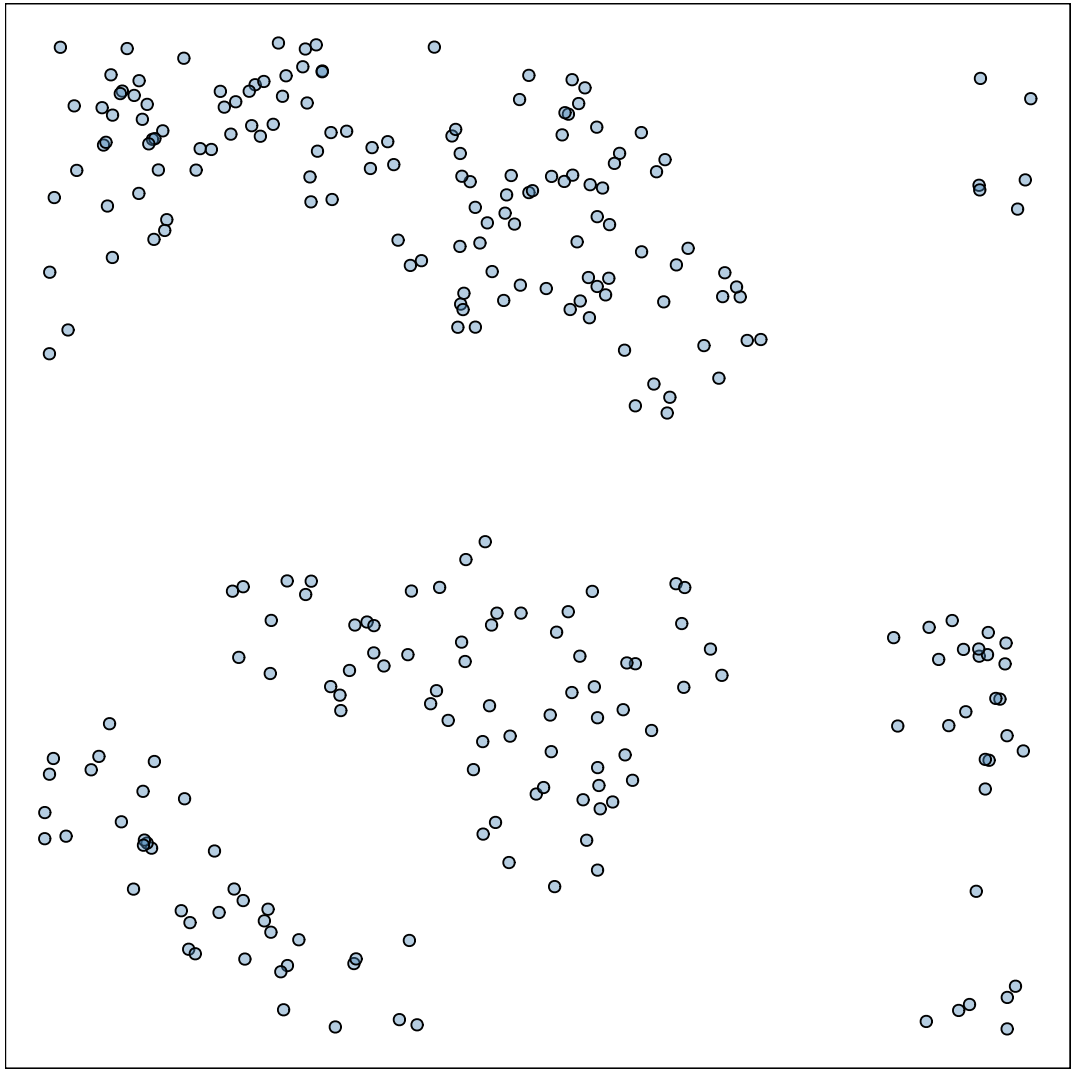
Clustered Point Patterns

- Some points are closer than expected by chance.
  - Groups of closely-spaced points
  - Extra space between clusters
- Clustered patterns can result from:



Clustered Point Patterns

- Some points are closer than expected by chance.
  - Groups of closely-spaced points
  - Extra space between clusters
- Clustered patterns can result from:
  - Attractive processes
  - Parent-offspring processes
  - Distribution of resources in the environment.
  - Others?



Pattern and Process: Point pattern analysis of *Juniperus occidentalis*

Spatial Patterns on the Sagebrush Steppe/Western Juniper Ecotone. Spatial pattern at 3 scales:

Small distances:

- Fewer neighbors than expected at distances less than 15 meters
- Inhibition: competition or light and water

Medium distances:

- More neighbors than expected by chance at 30-60 meters.
- More medium/large pairs between 50-70 meters
  - Fewer medium/small or large/small pairs
  - Short-distance seed dispersal by berry-eating birds:

Large distances:

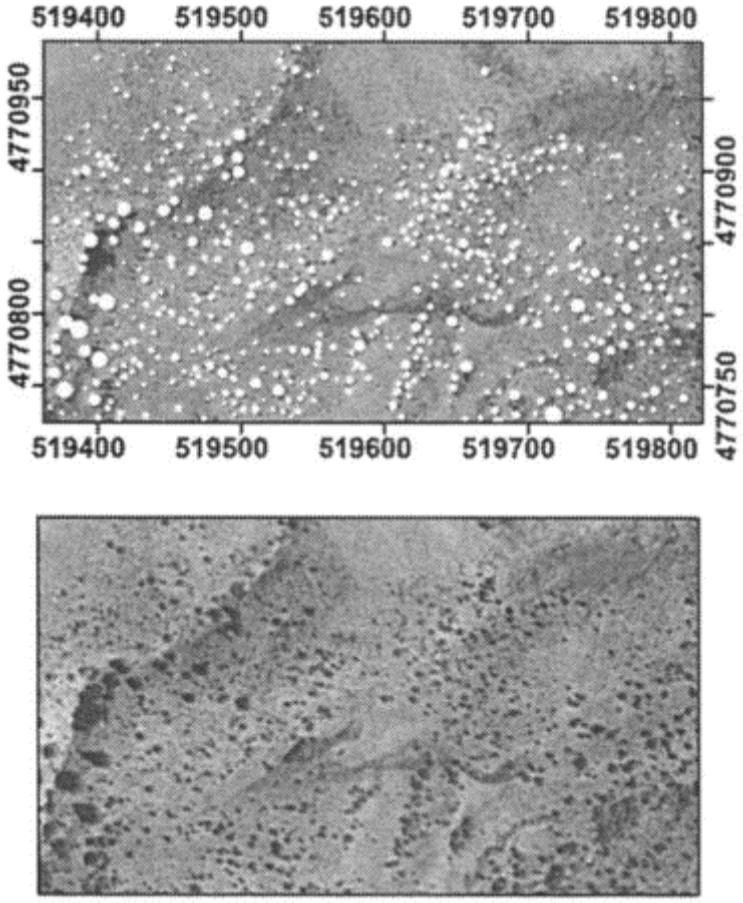


Figure 1 from Strand, E.K., Robinson, A.P., and Bunting, S.C. (2007).

Pattern and Process: Point pattern analysis of *Juniperus occidentalis*

Spatial Patterns on the Sagebrush Steppe/Western Juniper Ecotone. Spatial pattern at 3 scales:

Small distances:

- Fewer neighbors than expected at distances less than 15 meters
- Inhibition: competition or light and water

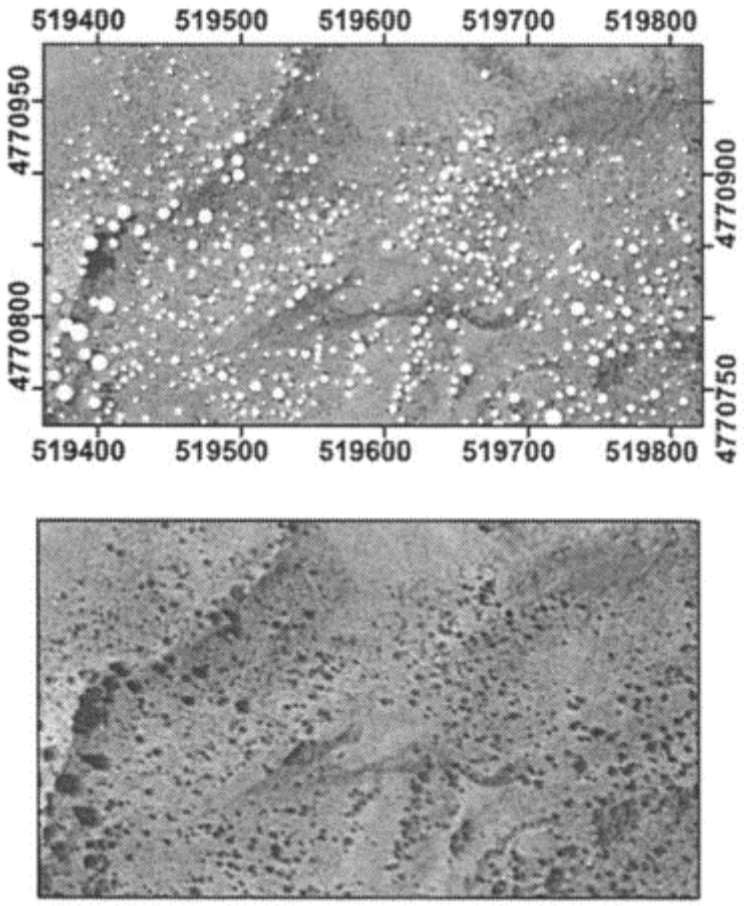


Figure 1 from Strand, E.K., Robinson, A.P., and Bunting, S.C. (2007).

Pattern and Process: Point pattern analysis of *Juniperus occidentalis*

Spatial Patterns on the Sagebrush Steppe/Western Juniper Ecotone. Spatial pattern at 3 scales:

Medium distances:

- More neighbors than expected by chance at 30-60 meters.
- More medium/large pairs between 50-70 meters
    - Fewer medium/small or large/small pairs
- Seed dispersal by berry-eating birds.

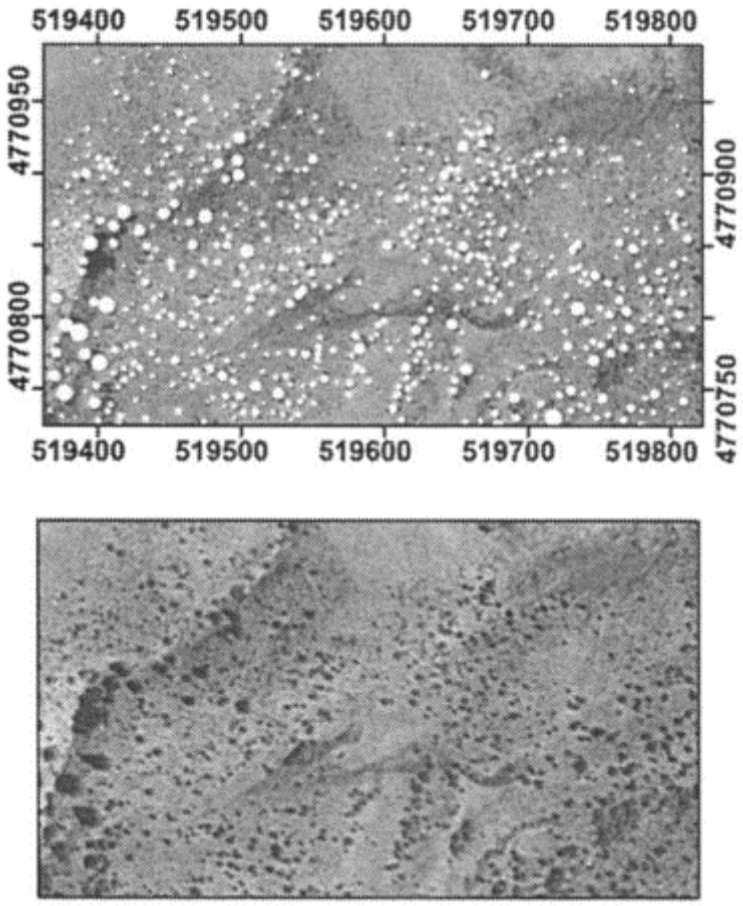


Figure 1 from Strand, E.K., Robinson, A.P., and Bunting, S.C. (2007).

Pattern and Process: Point pattern analysis of *Juniperus occidentalis*

Spatial Patterns on the Sagebrush Steppe/Western Juniper Ecotone. Spatial pattern at 3 scales:

Large distances:

- **N**o dependence at distances greater than 70 meters.
- Complete Spatial Randomness (CSR)

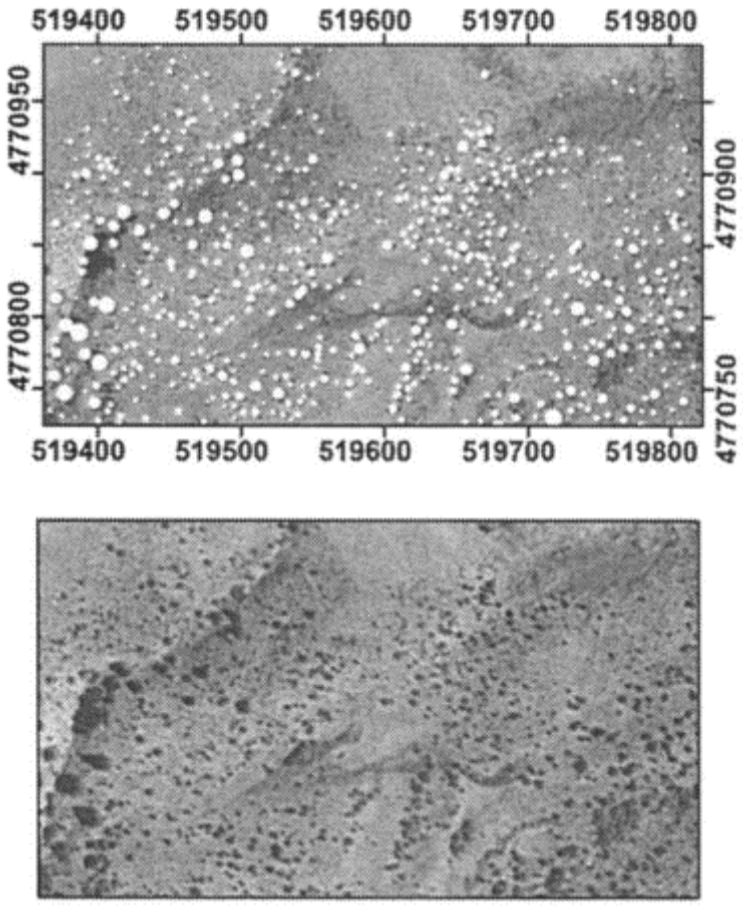


Figure 1 from Strand, E.K., Robinson, A.P., and Bunting, S.C. (2007).

How might we quantify pattern?

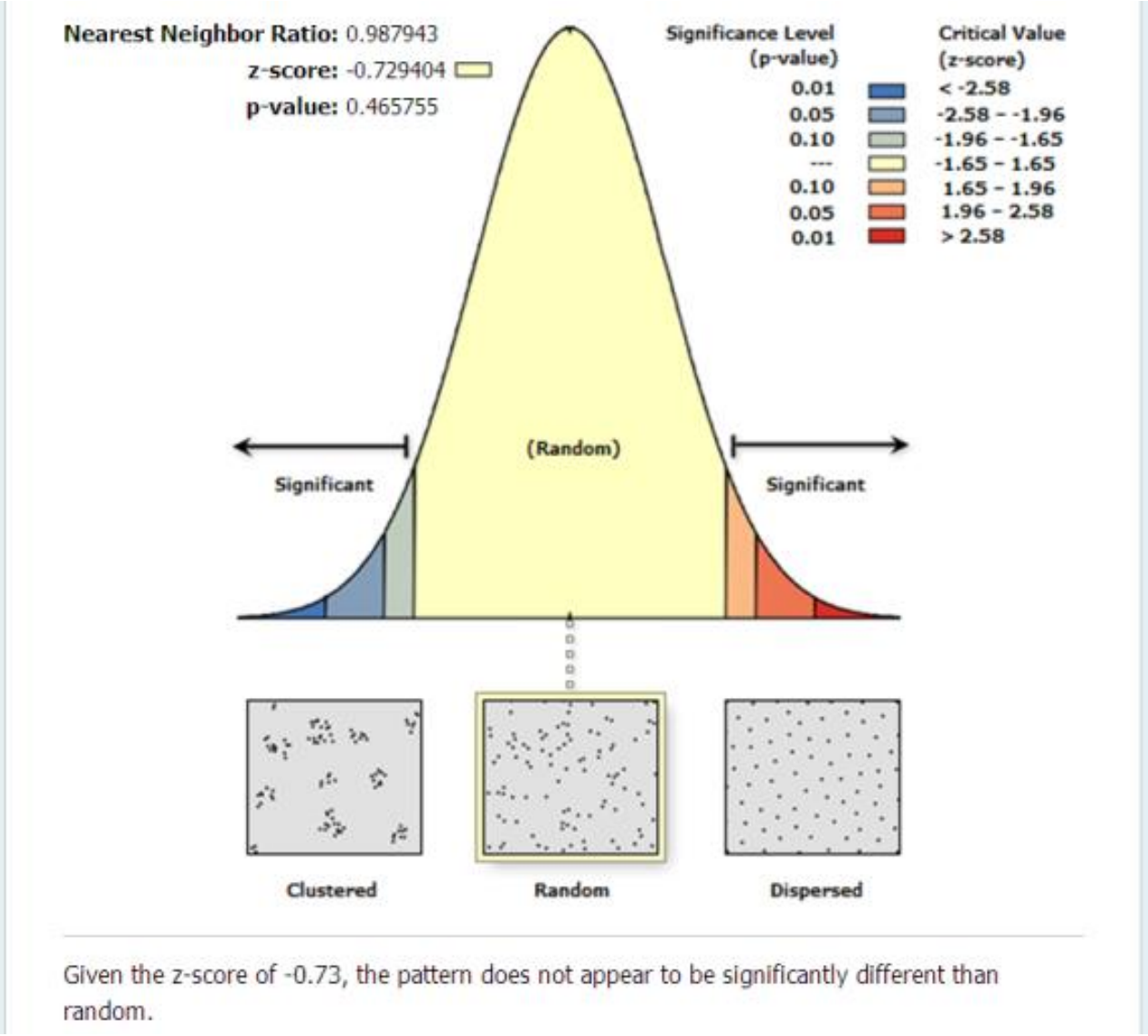
Number of points in a circle of radius r?

Average distance to nearest point?

Average distance to k nearest points?

Cluster analysis

Are points closer, or further apart, from each other than expected by chance?



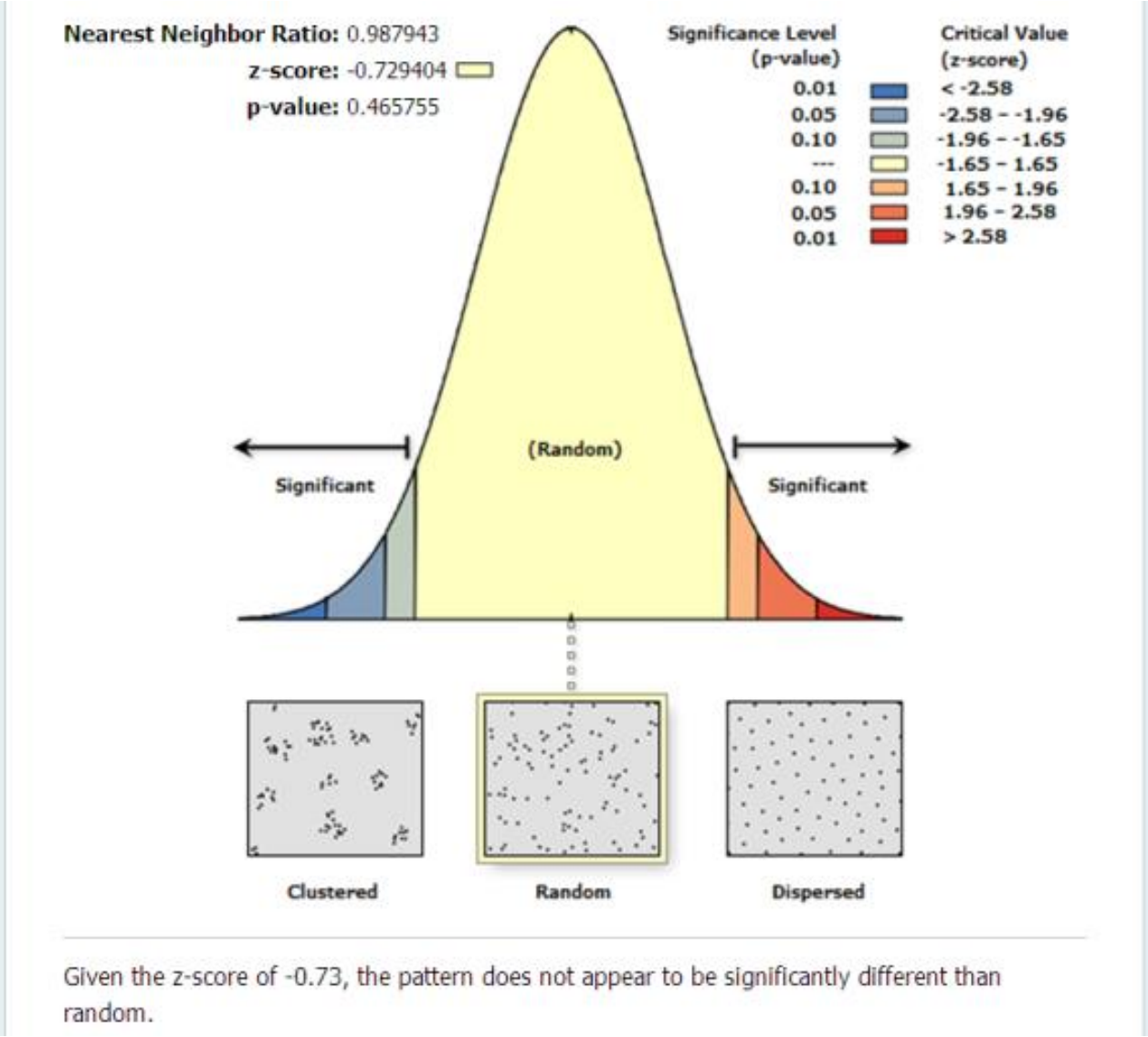
Cluster analysis

Cluster analysis quantifies **pattern**.

What about the **process** that leads to pattern?

Multiple processes can lead to the same pattern.

Can we infer process from pattern?



Spatial Point Data: Sampling

How do you choose a sampling strategy?

- Completely Random Sampling (CRS)?
- Convenience sampling?
- Transects?
- Stratified sampling?
- What are your constraints?
- Are your samples independent?

Where do point locations come from?

These data already exist, you don't really have control over sampling strategies.

But, what if you want to collect new samples?

Herbarium records

Where do point locations come from?

GPS collars

Herbarium records

Camera traps

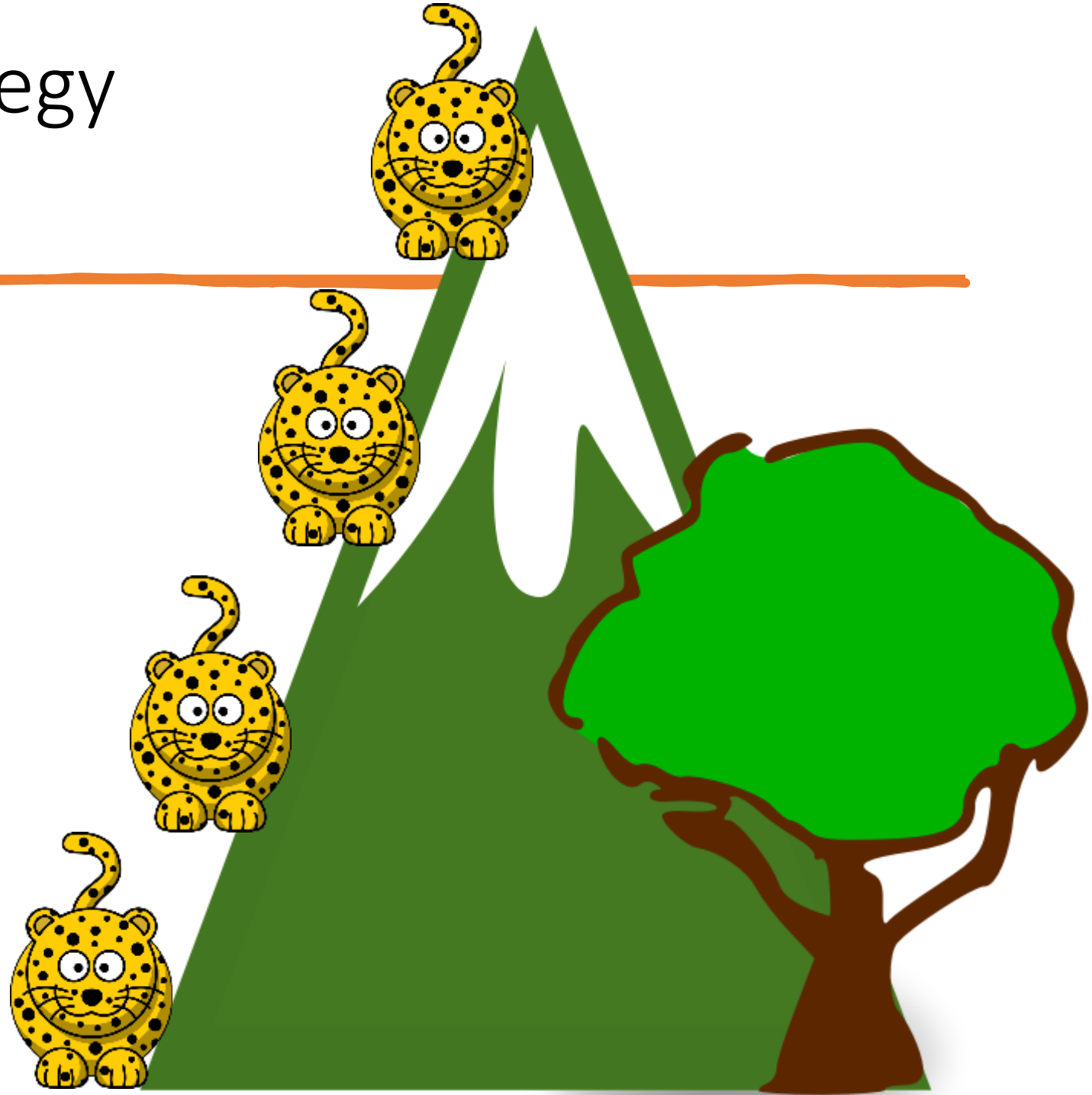
Point transects

Sampling Strategy

- What is my habitat?
- Available GIS layers:
  - Elevation
  - Water bodies
- Where should I place camera traps?

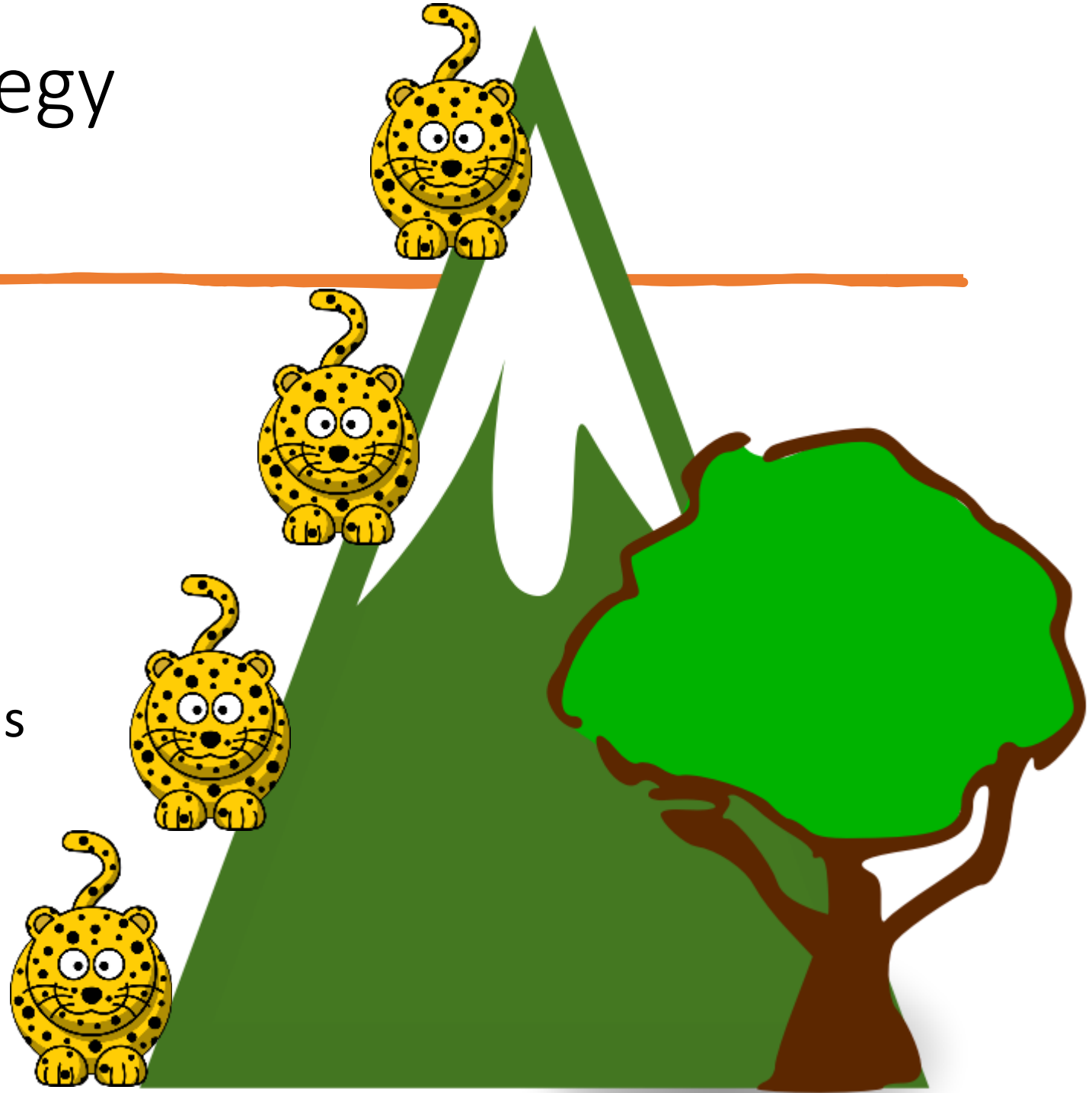
Sampling Strategy

- If you hypothesize a habitat preference, <u>stratify.</u>
  - <u>Why?</u>

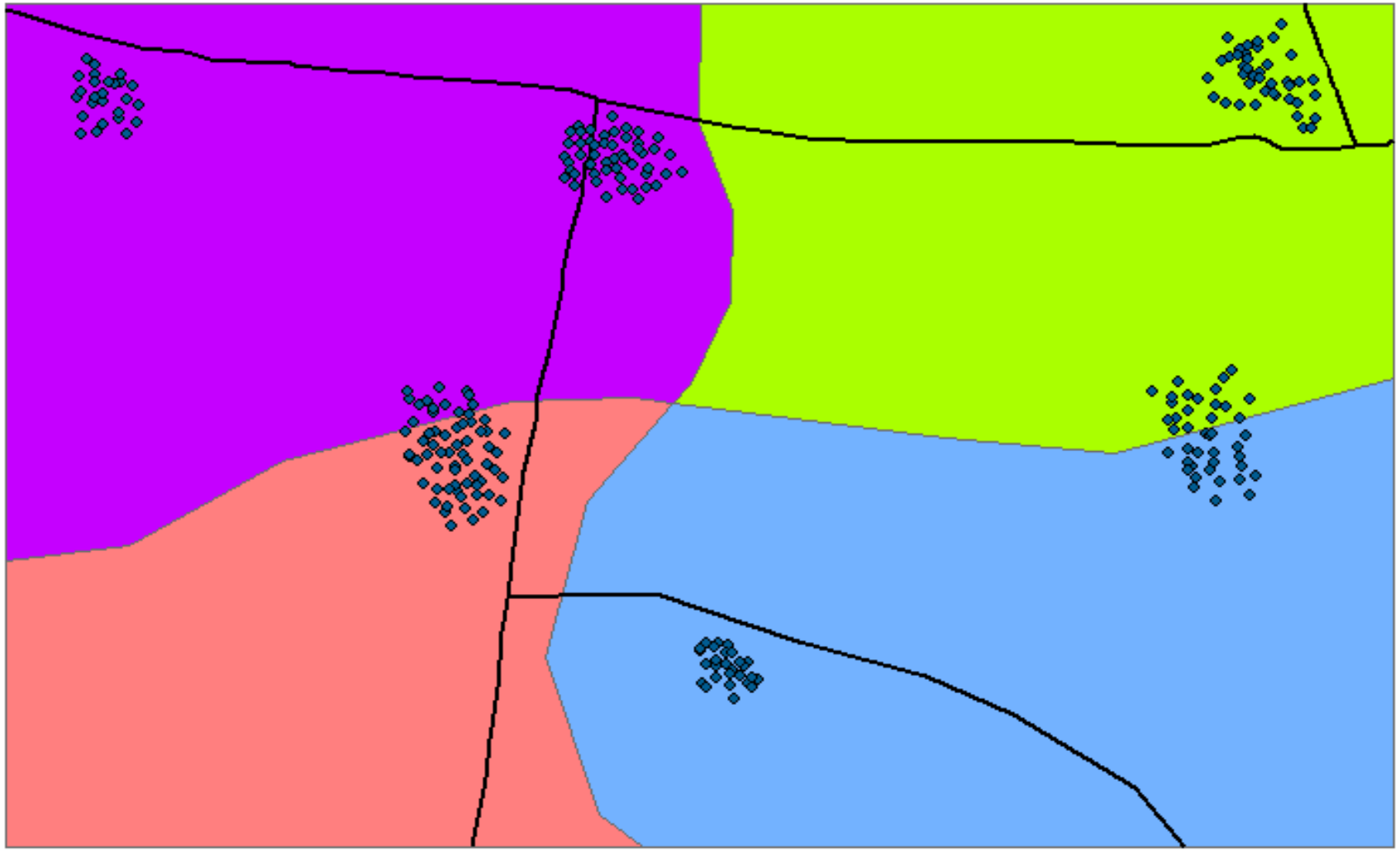


Sampling Strategy

- If you hypothesize a habitat preference, stratify
- But, try to hold other habitat variables constant
- Use GIS to help select locations that fit your criteria

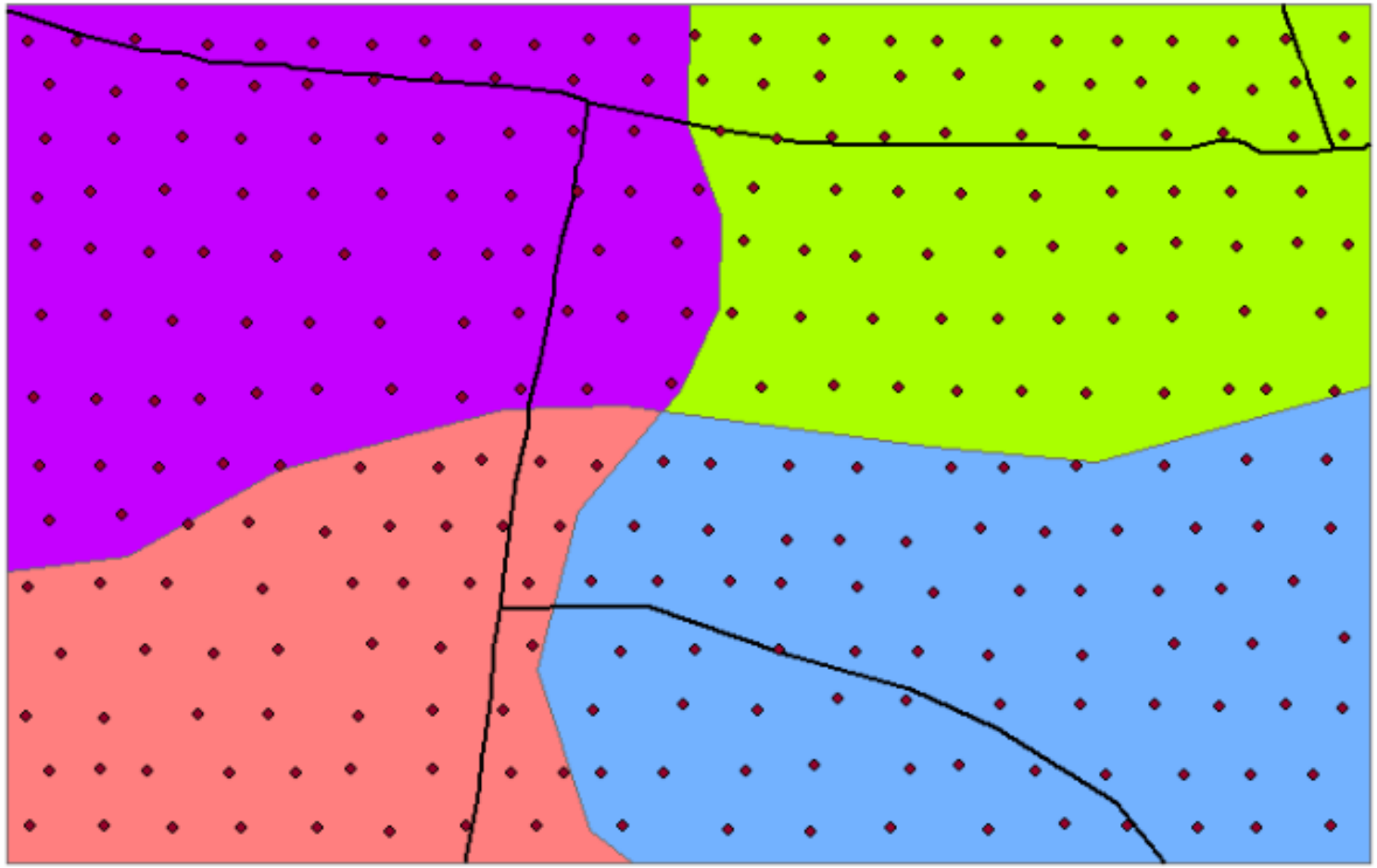


Is this sample size really 250 points?



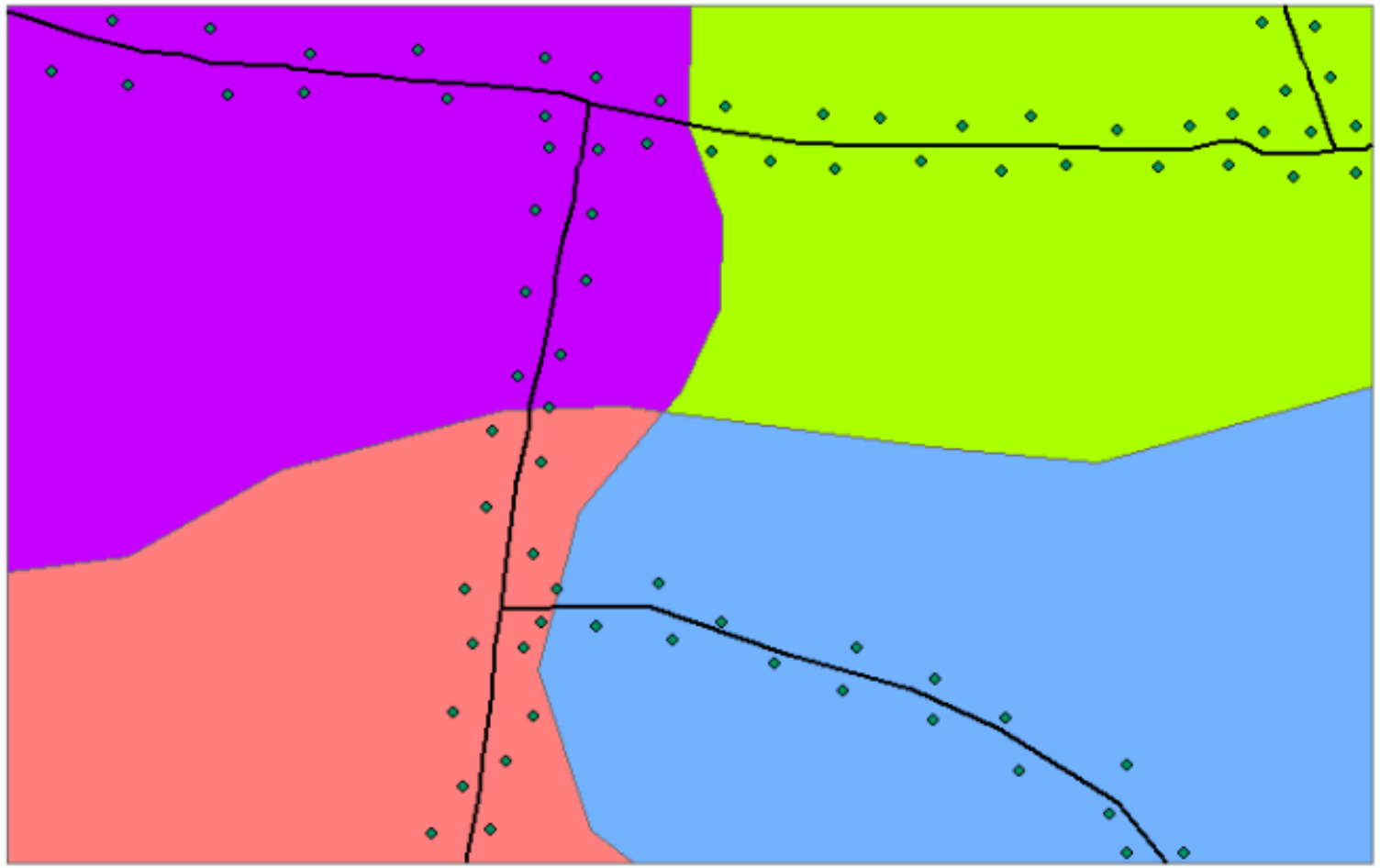
How do we collect points in the field?

- N = 250
- What is the sampling Strategy?
- Any potential issues?



How do we collect points in the field?

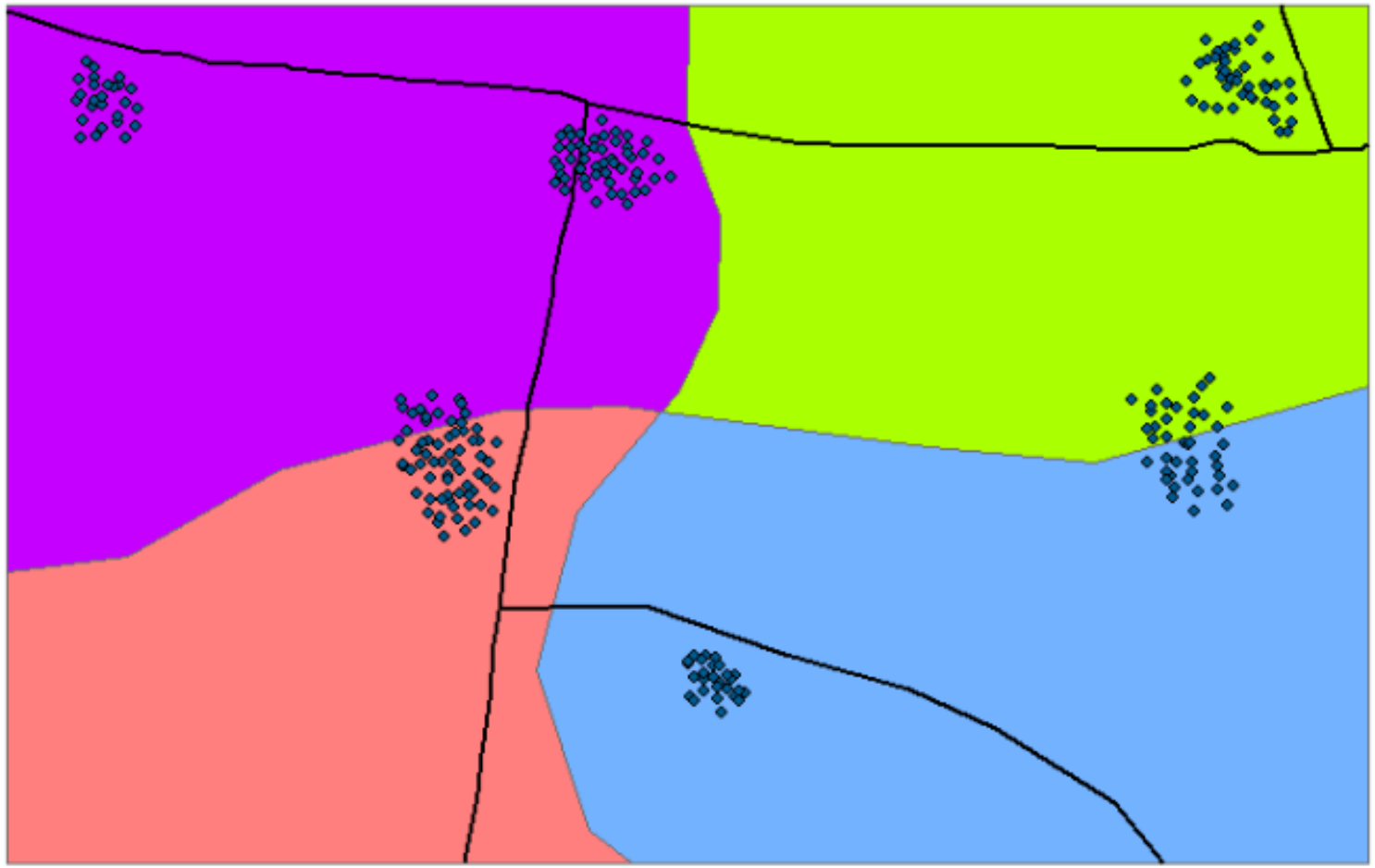
N = 75



How do we collect points in the field?

N = 250

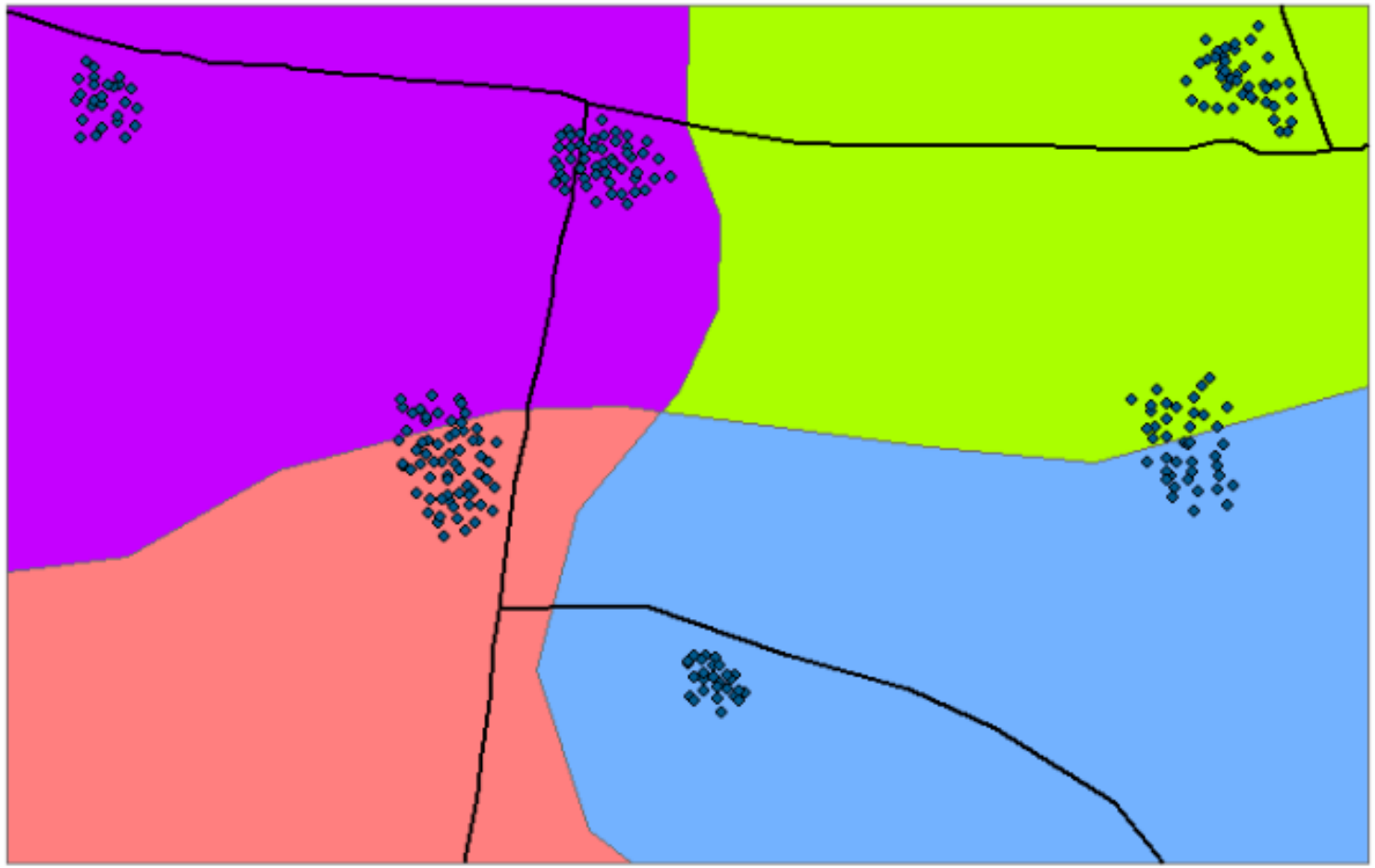
- What is the sampling Strategy?
- Any potential issues?



How do we collect points in the field?

Sometimes there's less data in the data...

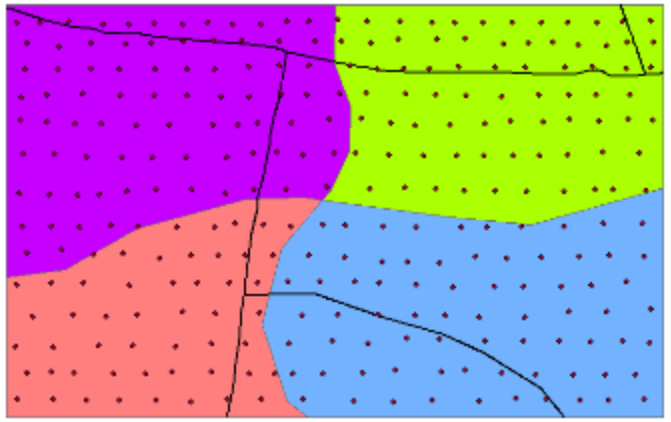
Non-independence means less information (and more difficult statistical inference).

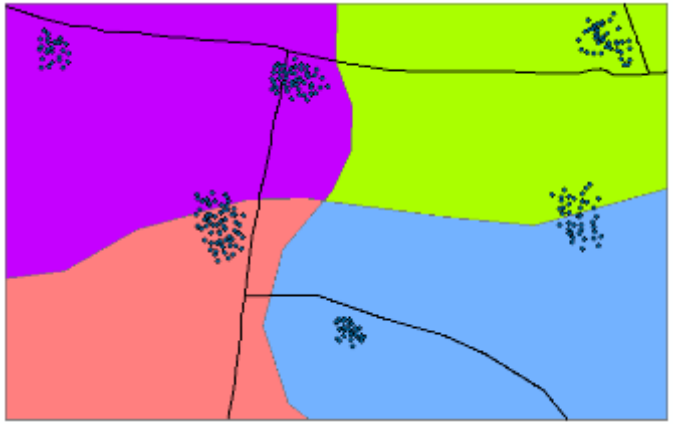


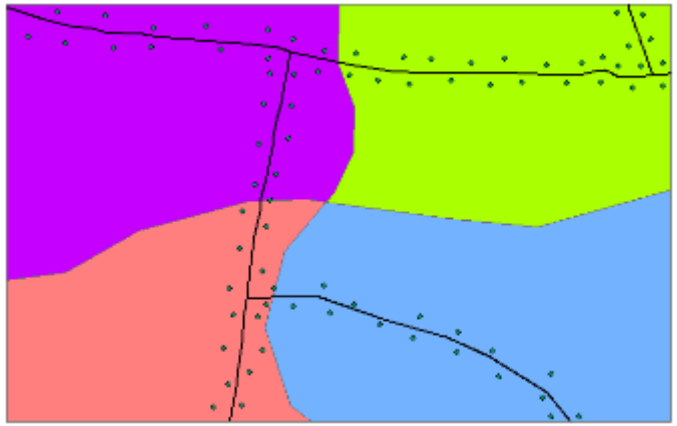
How do we collect points in the field?

Field sampling is always a tradeoff between:

- Effort (and resources)
- Potentially biased results
- Information content







Spatial Autocorrelation

Tobler’s First Law of Geography

- "everything is related to everything else, but near things are more related than distant things.”
  - Waldo Tobler, 1970.
- What is the second law?

Examples:
Positive Spatial Autocorrelation

- Parent-offspring processes
- Land cover (vegetation types)
- Topography
- Annual rainfall
- Income level

Negative Spatial Autocorrelation?

http://gisgeography.com/spatial-autocorrelation-moran-i-gis/

Negative Spatial Autocorrelation Example

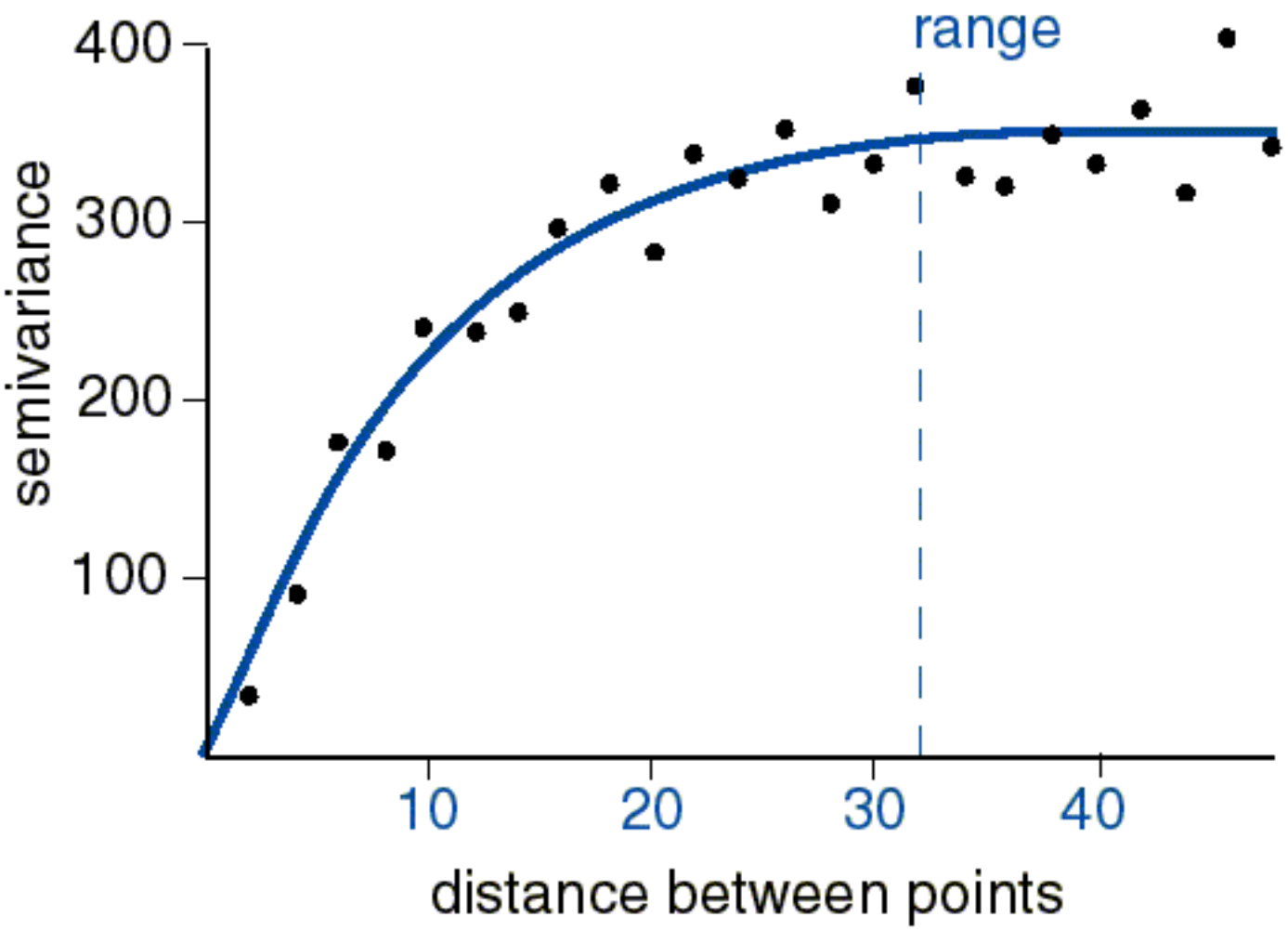
Negative spatial autocorrelation occurs when Moran's I is near -1. A checkerboard is an example where Moran's I is -1 because **dissimilar values are next to each other**. A value of 0 for Moran's I typically indicates no autocorrelation.

Checkerboard Pattern: Spatial Autocorrelation

Using the spatial autocorrelation tool in ArcGIS, the checkerboard pattern generates a Moran's index of -1.00 with a z-score of -7.59. (Remember that the z-score indicates the statistical significance given the number of features in the dataset). This checkerboard pattern has a less than 1% likelihood that it is the result of random choice.

Semivariogram – measures spatial autocorrelation

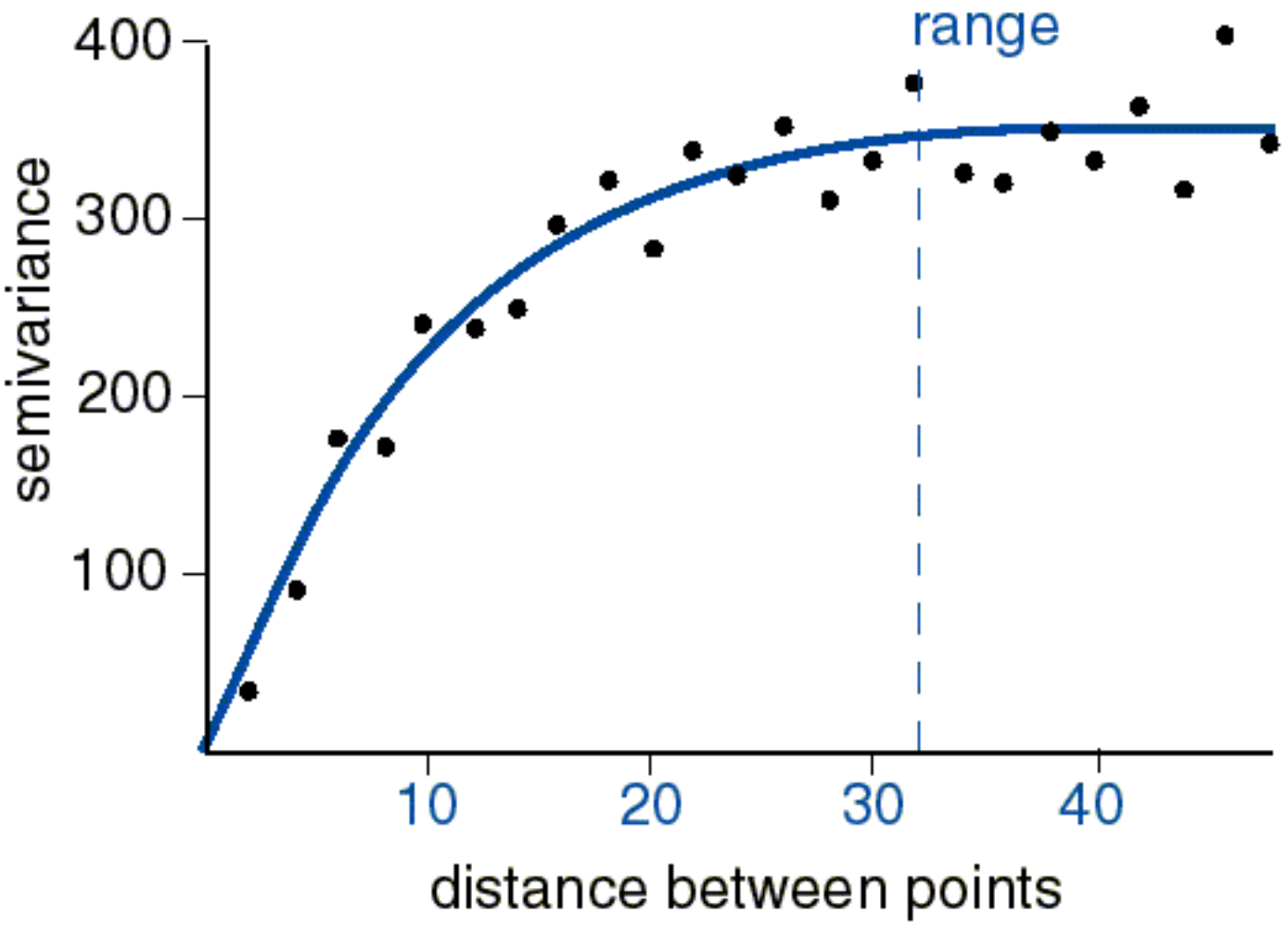
http://www.sciencedirect.com/science/article/pii/S016953479801533X



Range is a critical distance:

At shorter distances, values are more similar than you'd expect by chance. They are autocorrelated.

At greater distances, knowing the value of one observation doesn't give you information bout the second.

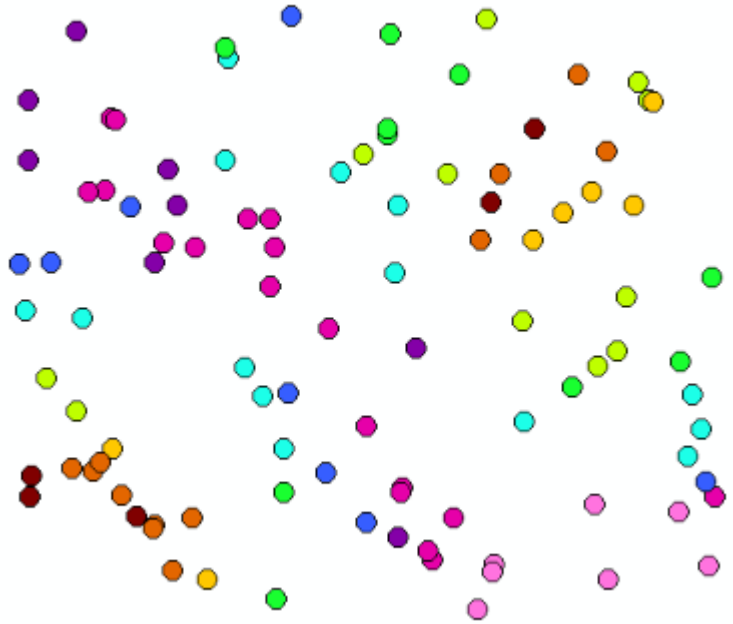


Interpolating Surfaces from Points

Intro to GIS – UMass Amherst – Michael F. Nelson

Interpolation: Why should you?

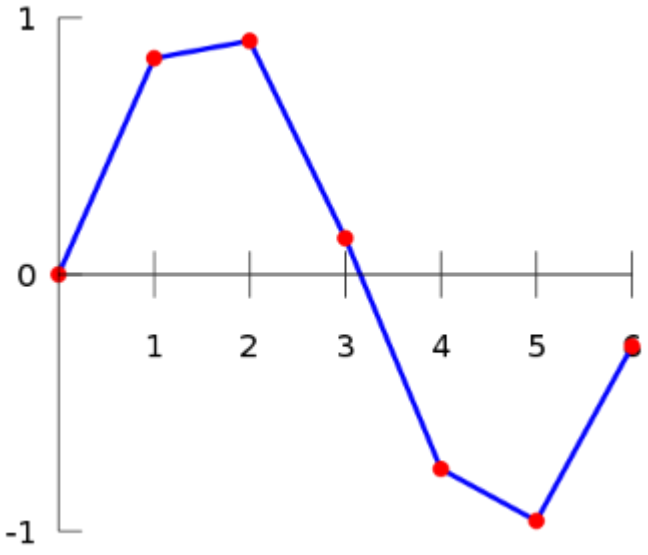
- Rain gages
- Temperature
- Population
- Groundwater level
- Ocean pollutants
- Crimes

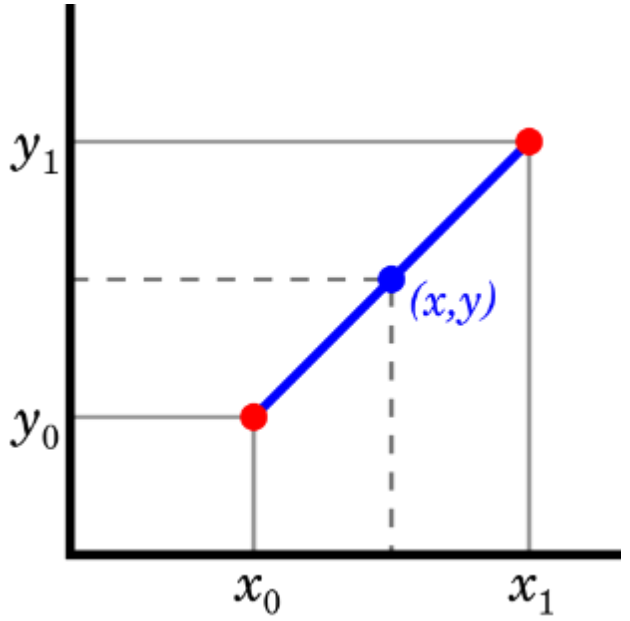


- Unfortunately, very few things come in continuous raster datasets!

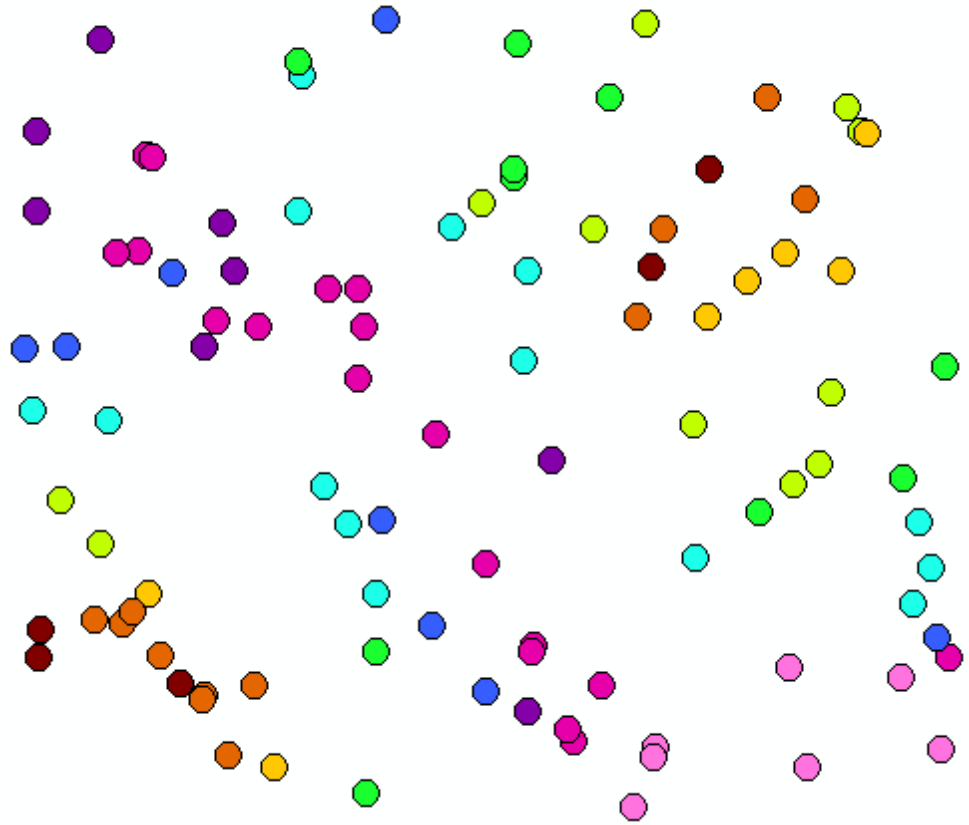
Interpolation

- At its core, interpolation creates new data from a range of a discrete set of known data.

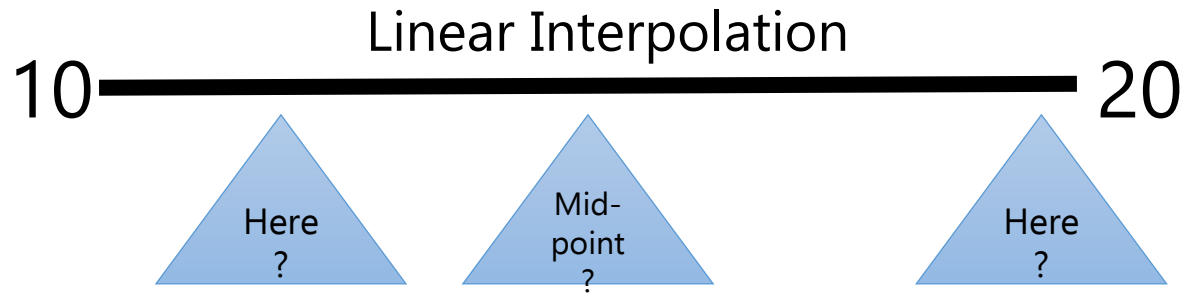


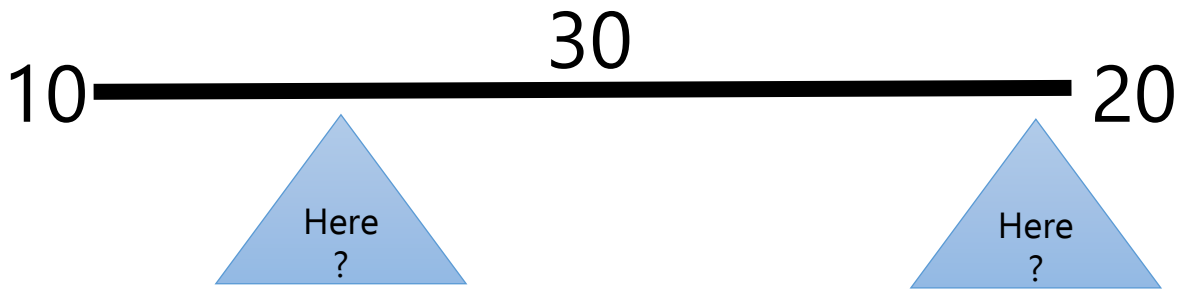


Sample Data Points



Practice!


Linear Interpolation
10
20
Here ?
Mid-point ?
Here ?


30
10
20
Here ?
Here ?

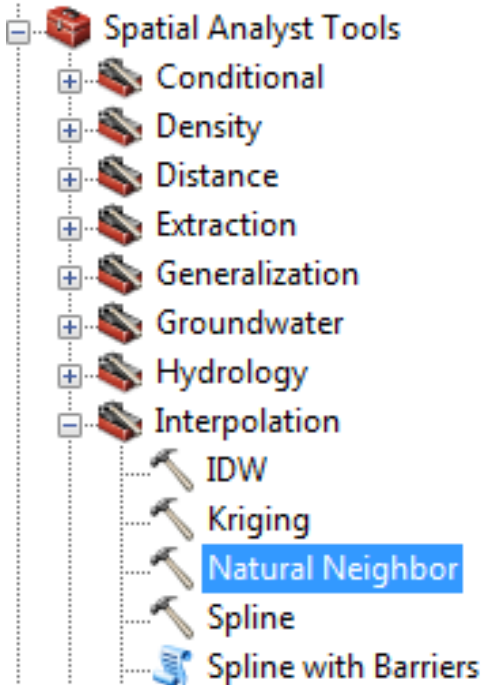
Four ways to interpolate

Natural Neighbor Triangulation

Inverse Distance Weighting

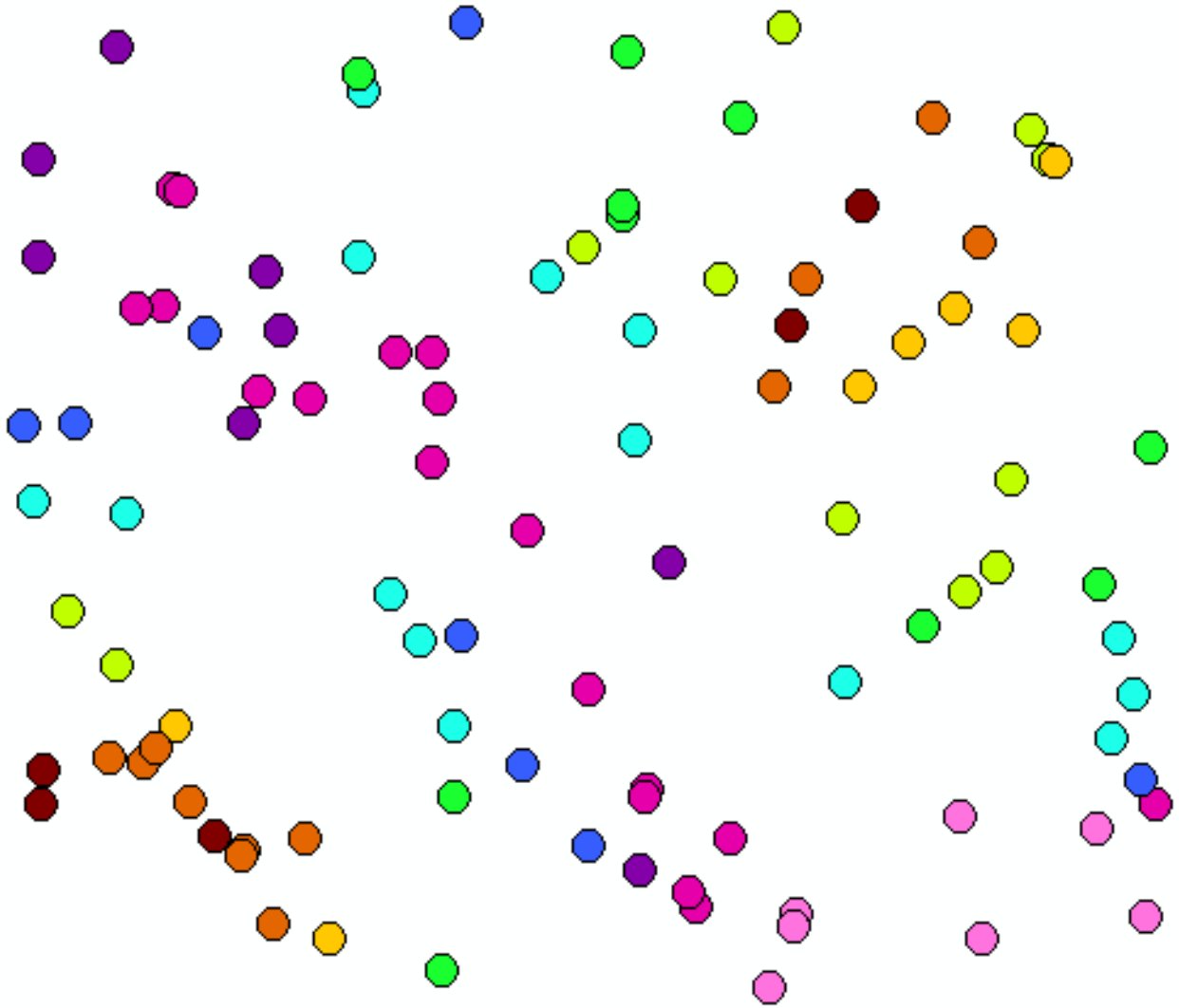
Splines

Kriging



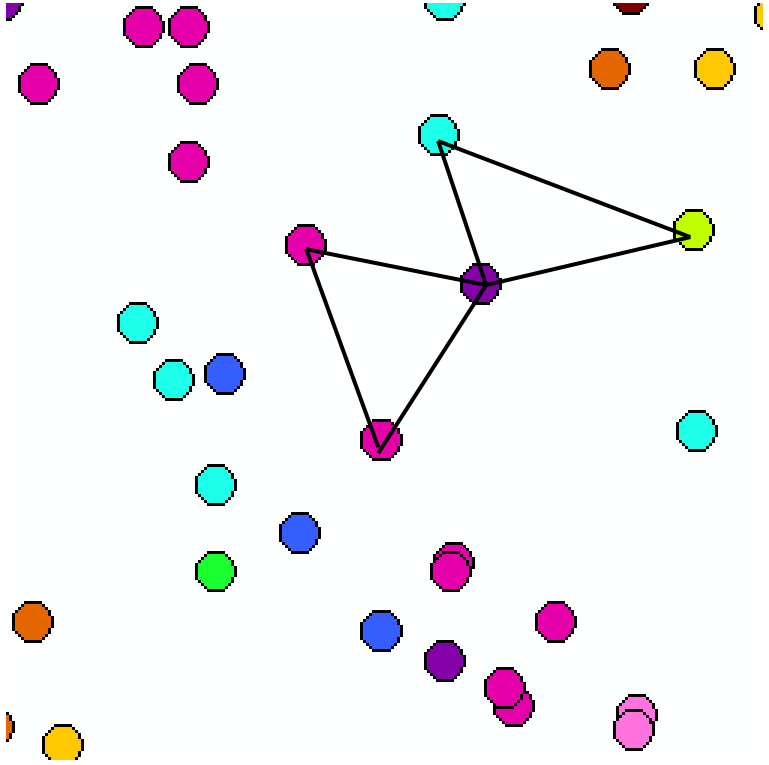
Sample Data Points

- The colors represent a numeric value



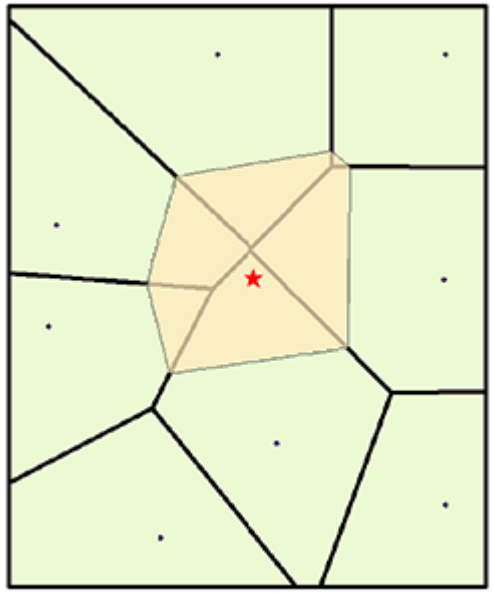
Natural Neighbor Triangulation

- Any data point can be interpolated based on a triangle of 3 known points.
- Also known as Delauney Triangulation



Natural Neighbor Triangulation

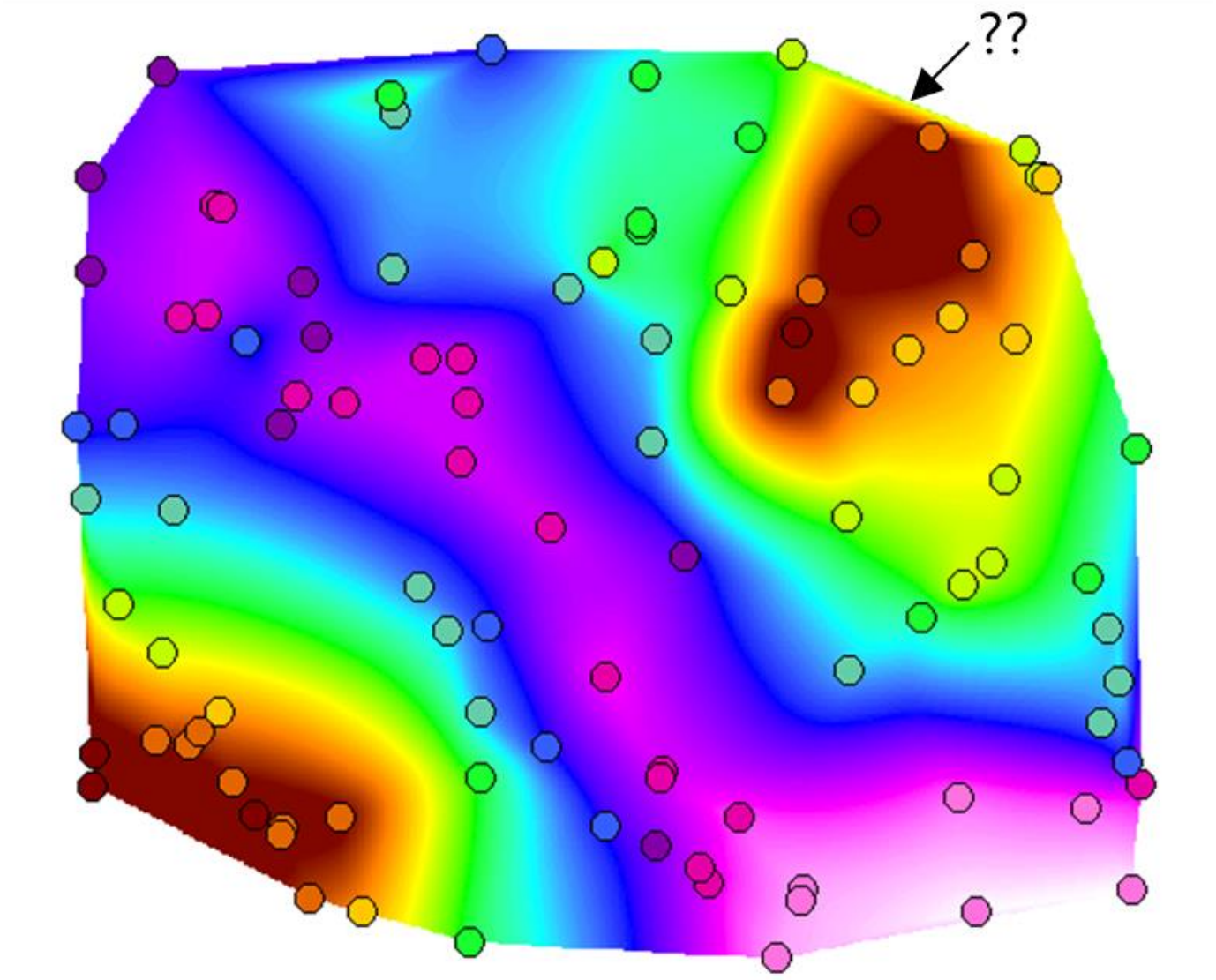
1. Voronoi tessellation from known points (black points)
2. New Voronoi polygon from unknown point (red star).
3. Value at unknown point is the weighted mean of the overlapping tessellation polygons.



https://pro.arcgis.com/en/pro-app/2.8/tool-reference/spatial-analyst/how-natural-neighbor-works.htm

Natural Neighbor Triangulation

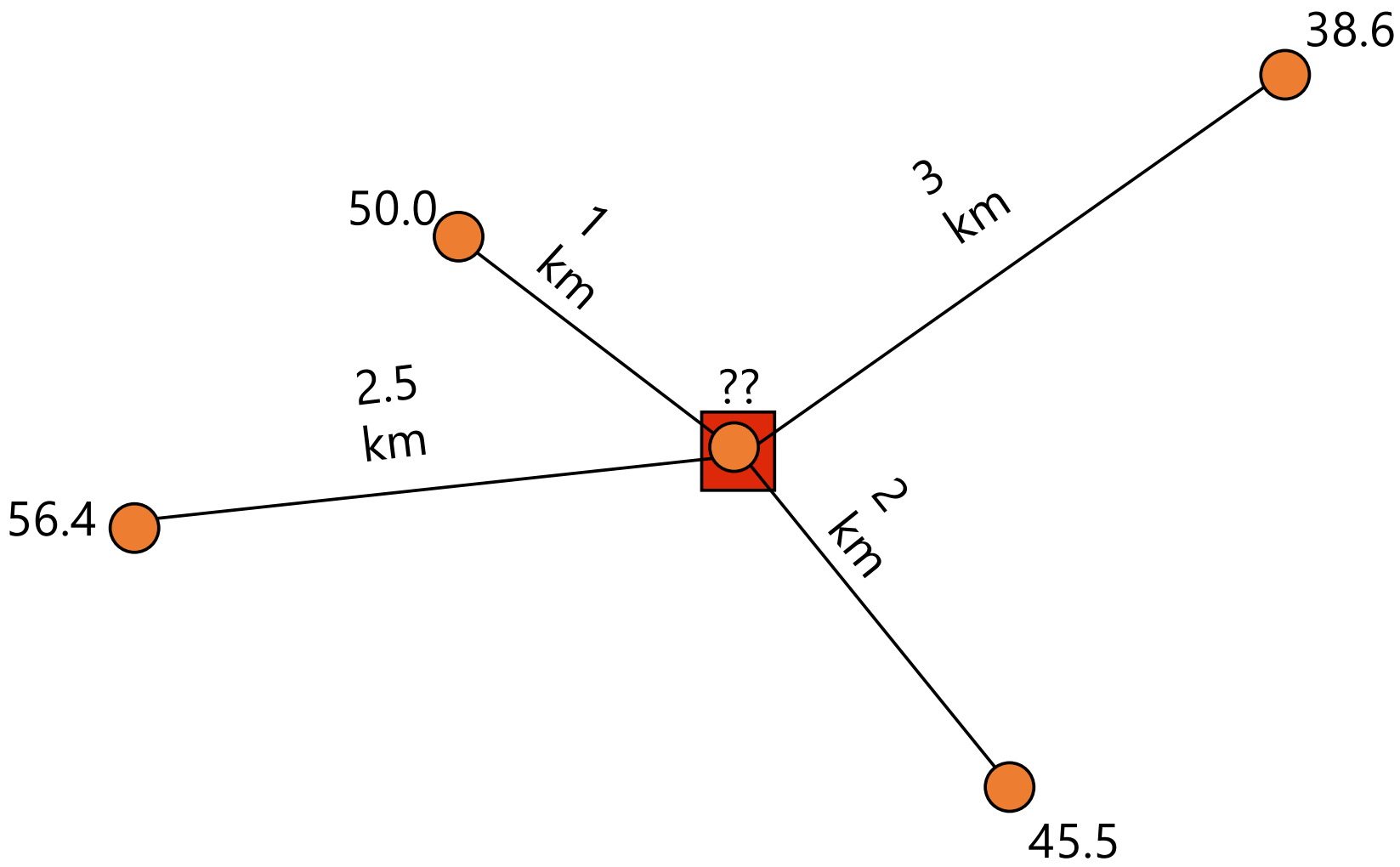
- Note: The raster output is not square!
- Why?
- Could we force it to be square?



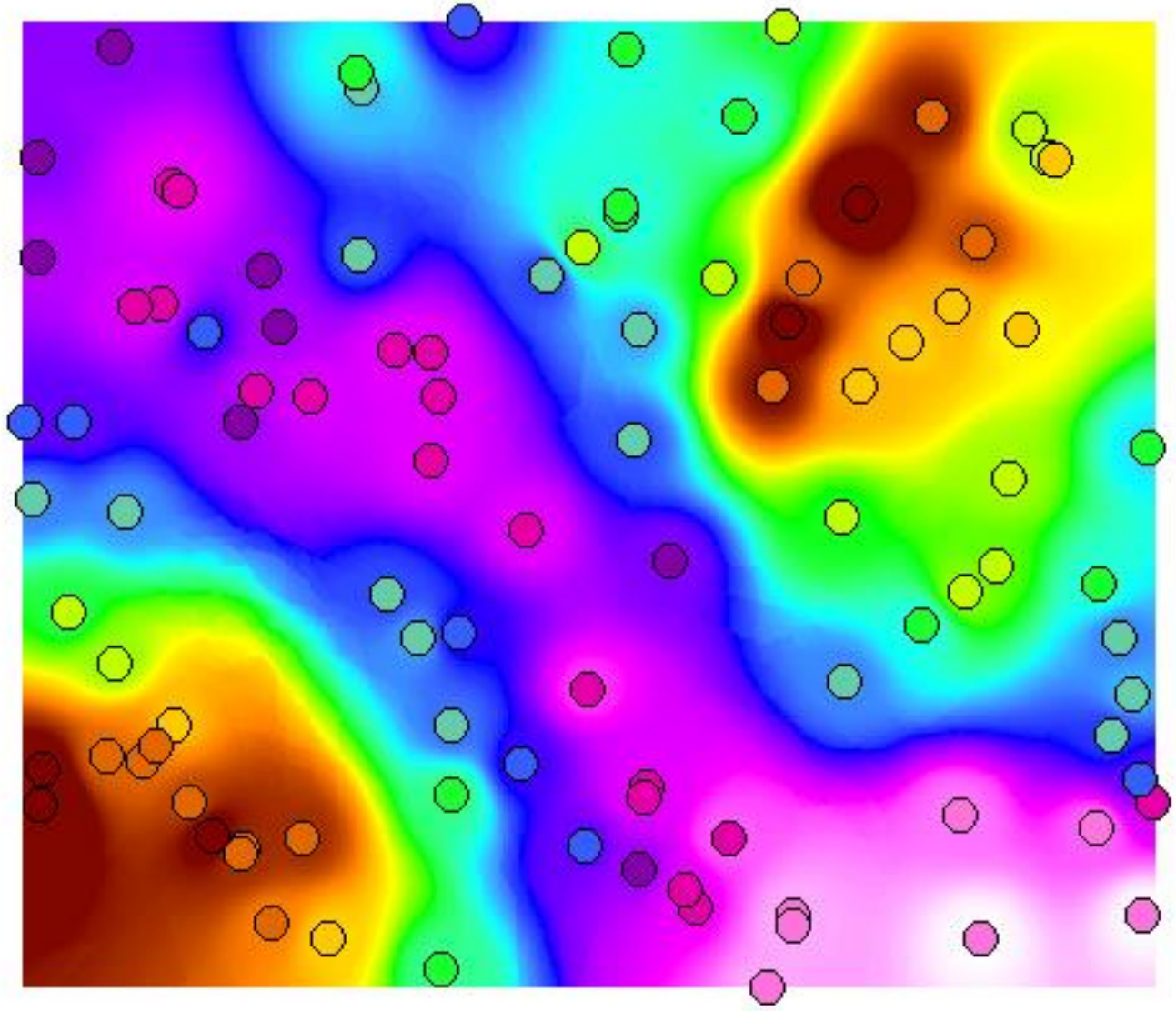
IDW: Inverse Distance Weighting

Greater local effect can produce ‘bullseye effects’

$W(d) = 1/d^p$



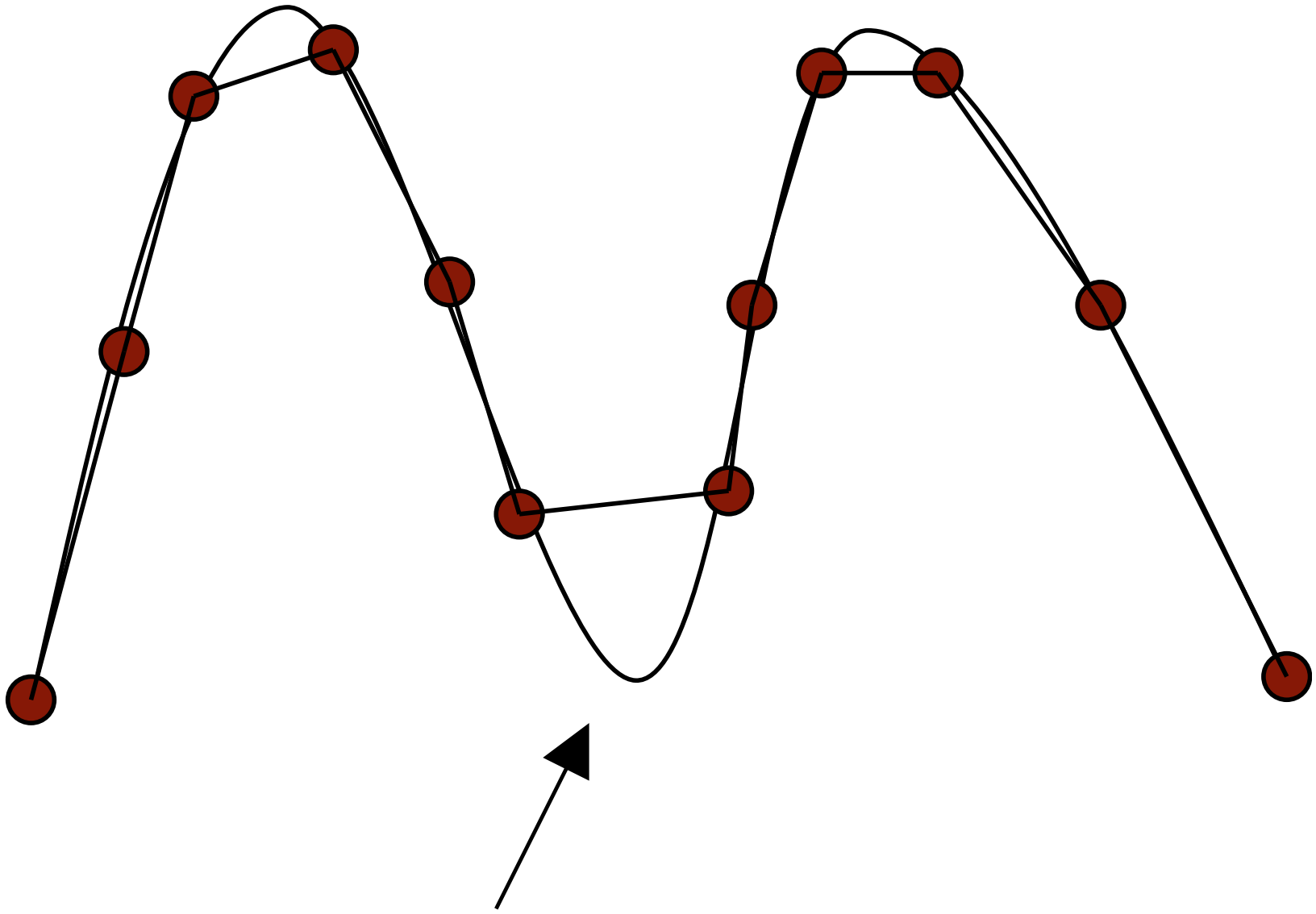
IDW Interpolation



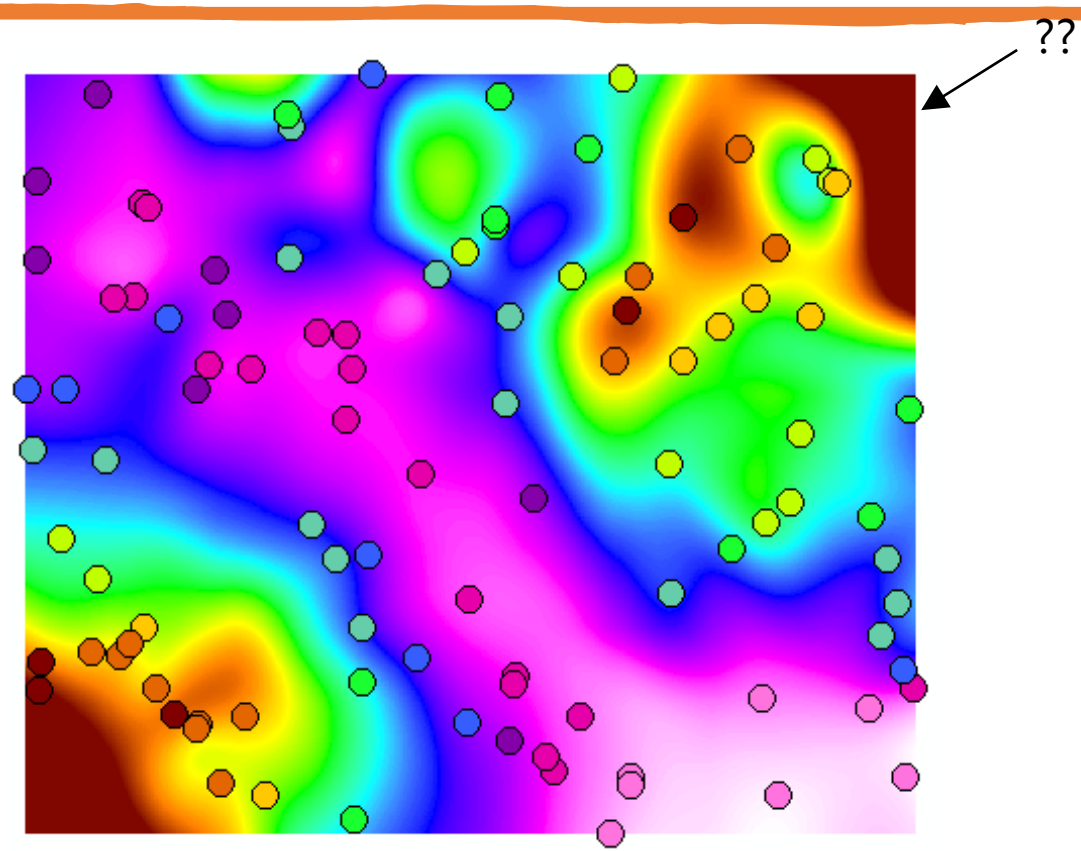
Spline Interpolation

Splines are smoothed curves fit through data points

Mathematically minimized curvature, but it can predict values above maxima and below minima



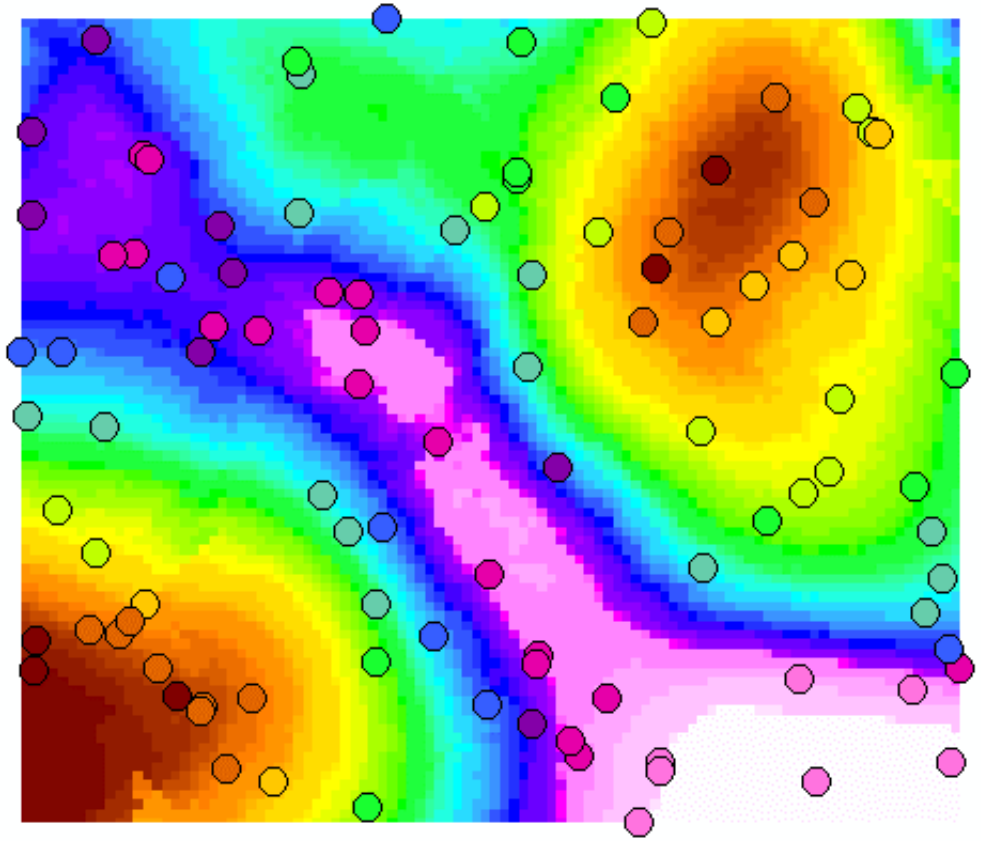
Spline Interpolation



Kriging Interpolation

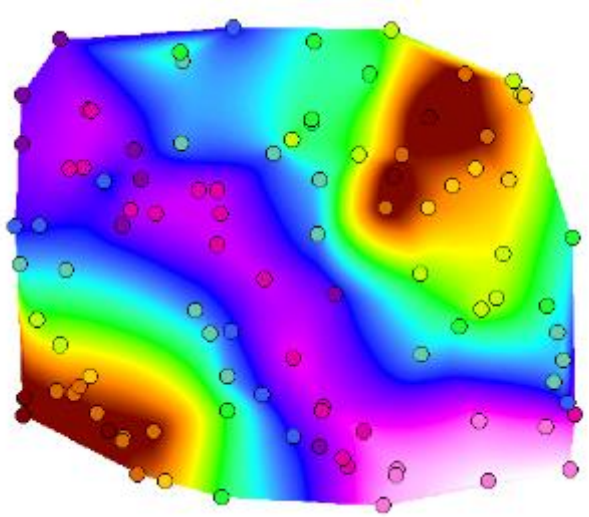
- Kriging is a statistical modeling technique. It includes deterministic and stochastic models.
- Combines distance weighting with knowledge of spatial correlation
- Includes measure of error on the final estimate
- More powerful than the other interpolations, but also requires more knowledge to implement well

Kriging Interpolation (Prediction)

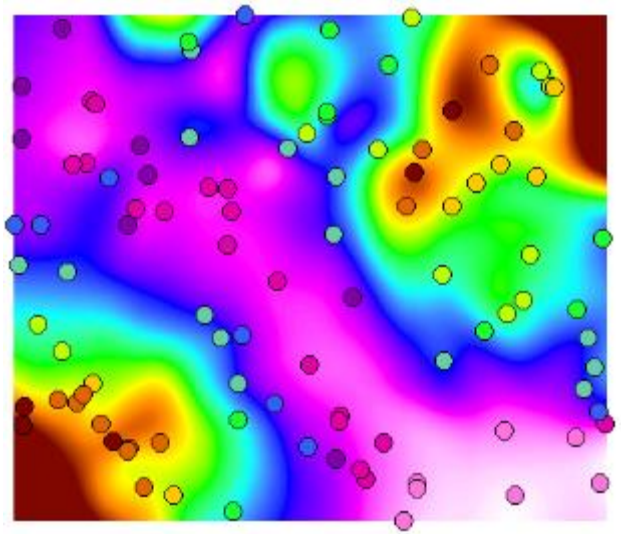


Interpolation Results

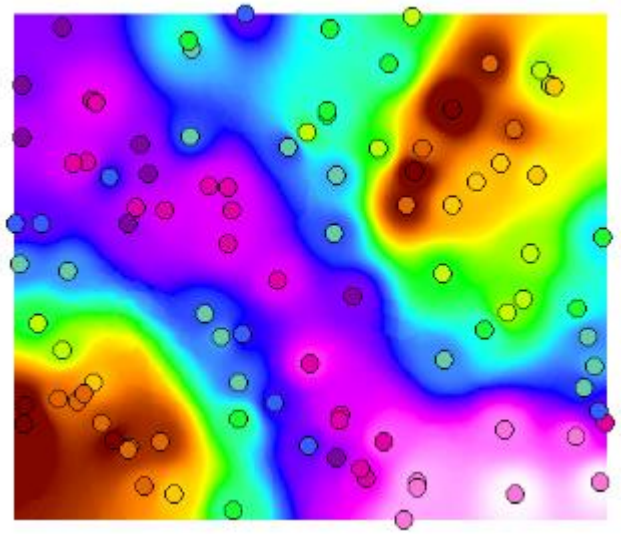
Natural Neighbor



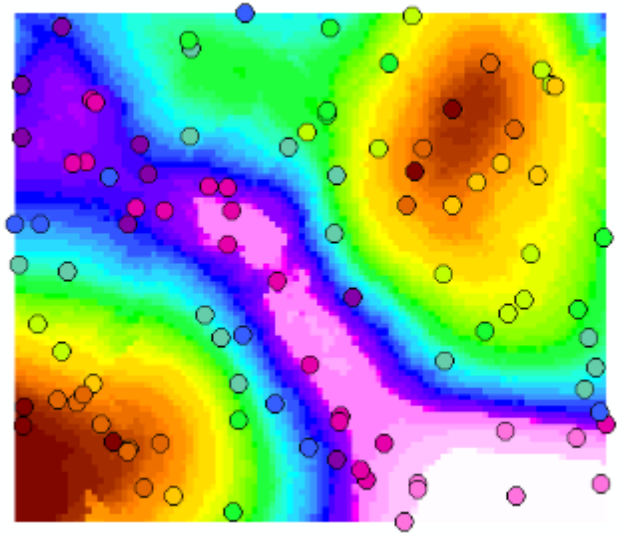
Splines



IDW



Kriging



Interpolation: Take home messages

- There are multiple ways to interpolate surfaces between data points – each with its own strengths and weaknesses
- Your choice of interpolation method will depend on your data, your modeling purpose, your judgment.
- The best interpolation in the world can’t make up for lack of data!!

Final Projects

Questions?