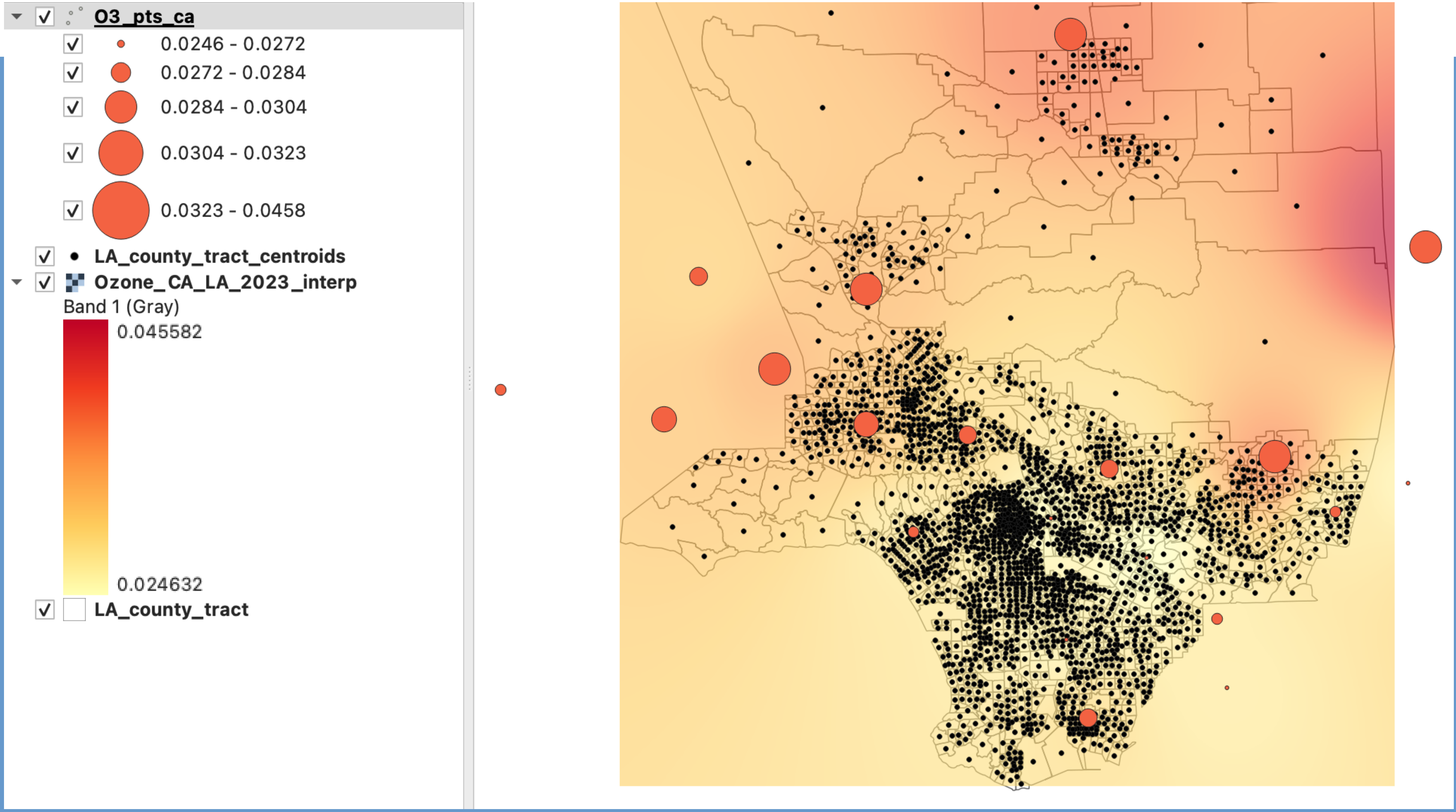
Spatial Interpolation

Lecture #20 | GEOG 510
GIS & Spatial Analysis in Public Health

Varun Goel

Outline

- Interpolation
  - Point-based methods
    - Nearest neighbor
    - Trend surface
    - IDW
    - Kriging (classical, EB)


O3_pts_ca
0.0246 - 0.0272
0.0272 - 0.0284
0.0284 - 0.0304
0.0304 - 0.0323
0.0323 - 0.0458
LA_county_tract_centroids
Ozone_CA_LA_2023_interp
Band 1 (Gray)
0.045582
0.024632
LA_county_tract

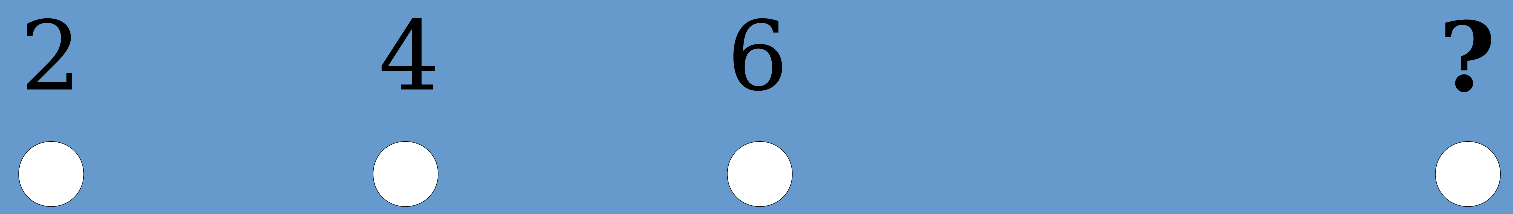
What is the missing value?

10 20 ? 40 50

What is the missing value?

10 20 **30** 40 50

What is the missing value?



What is the missing value?

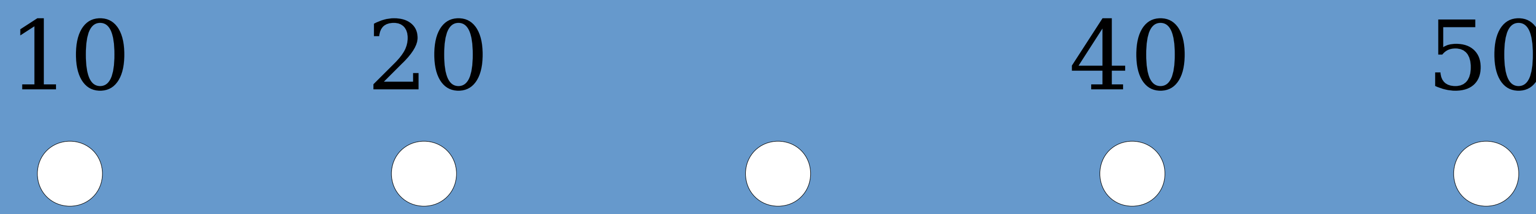
2 4 6 **10**

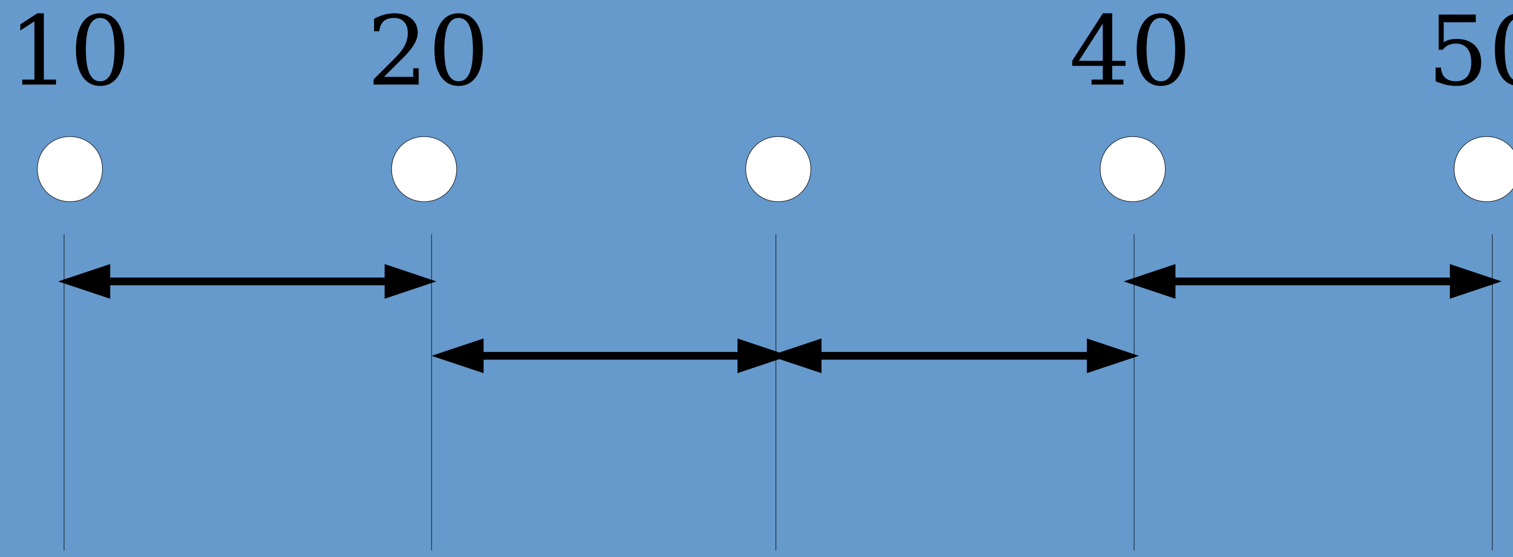
What is the missing value?

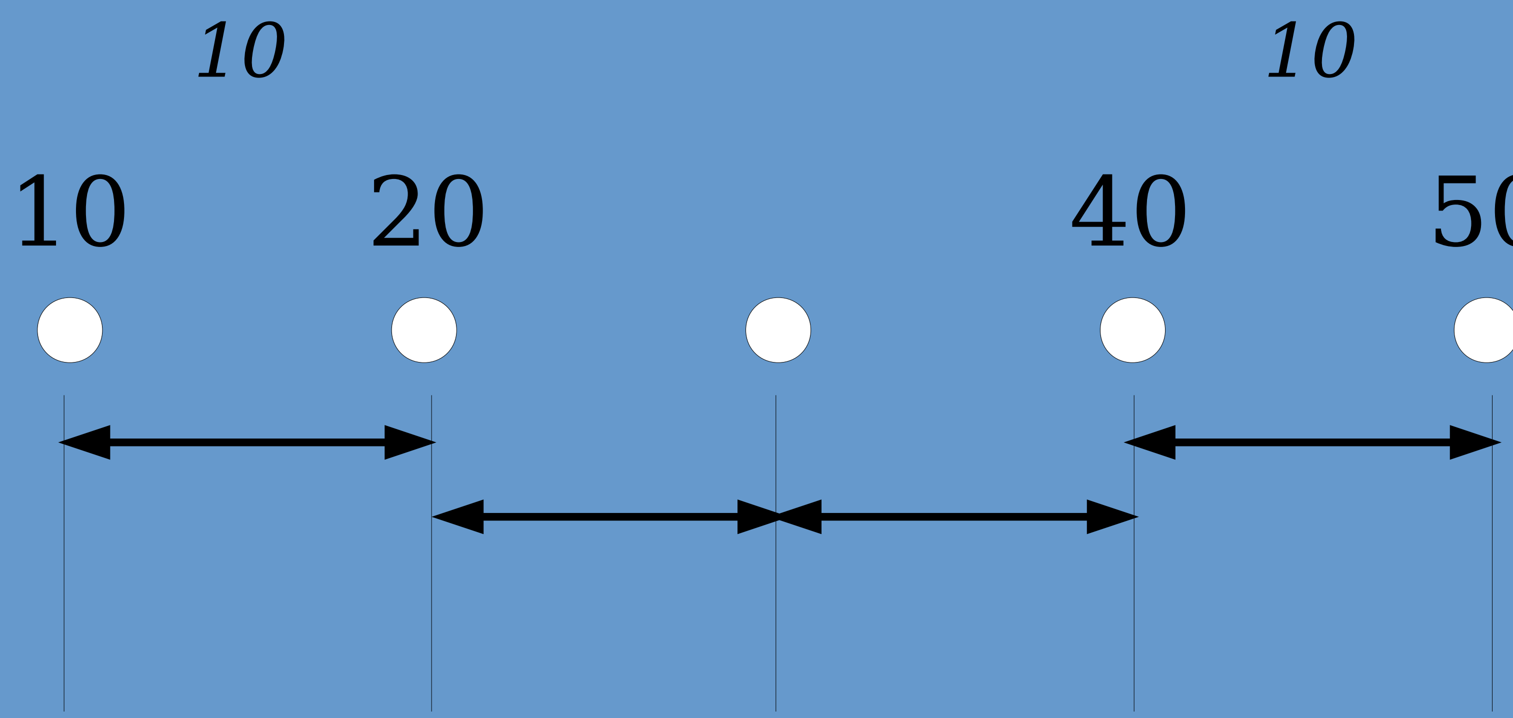
2 4 6 4 ?

What is the missing value?

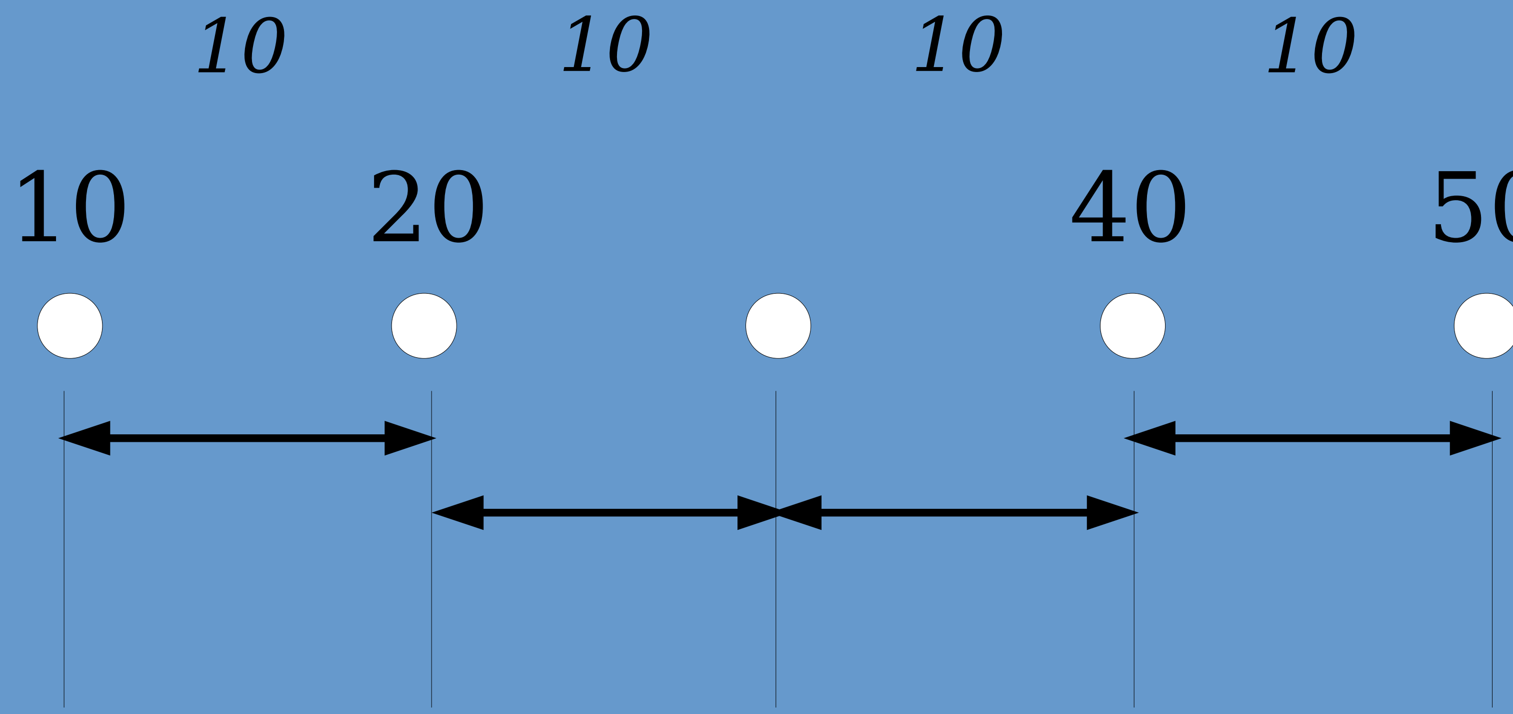
2 4 6 4 **2**

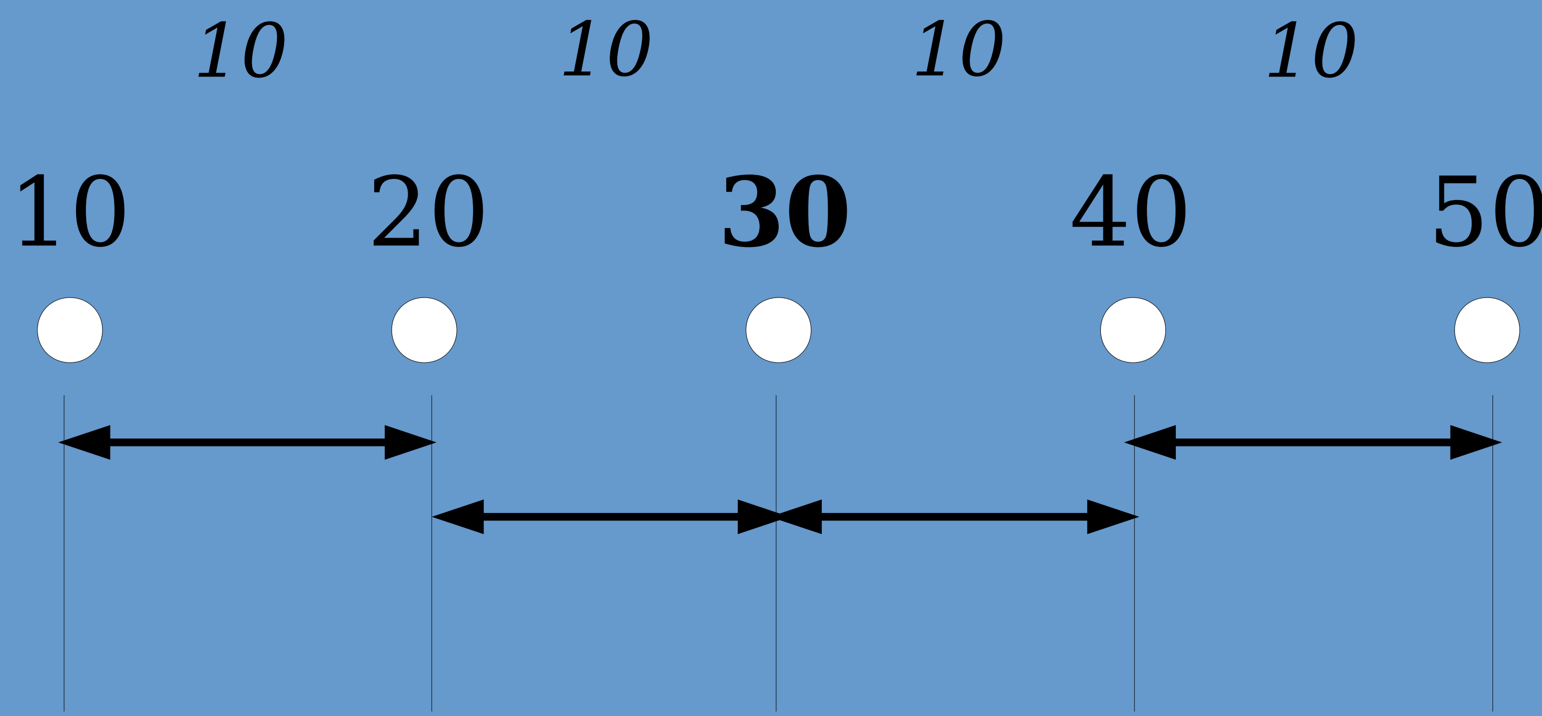

10
20
40
50


10
20
40
50

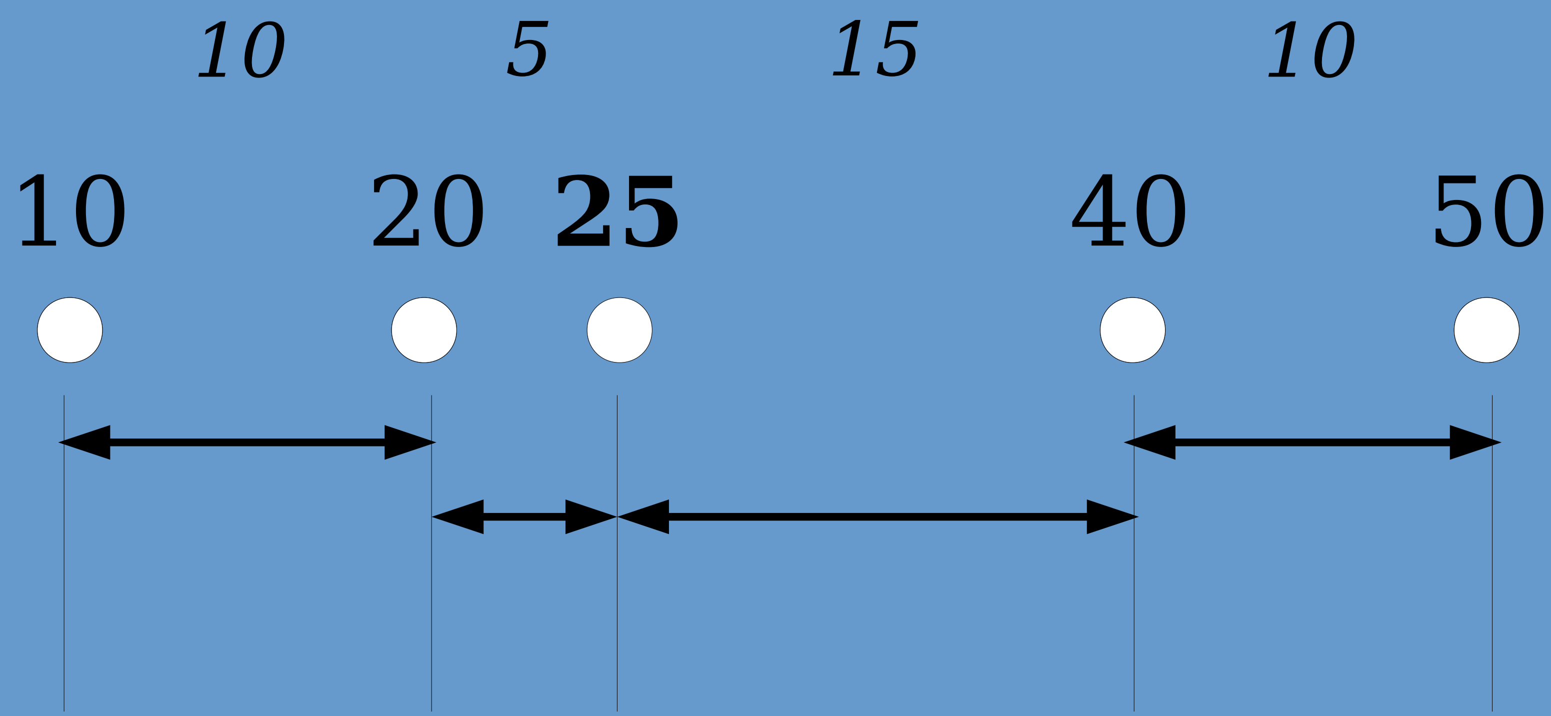


10
10
10
20
40
50

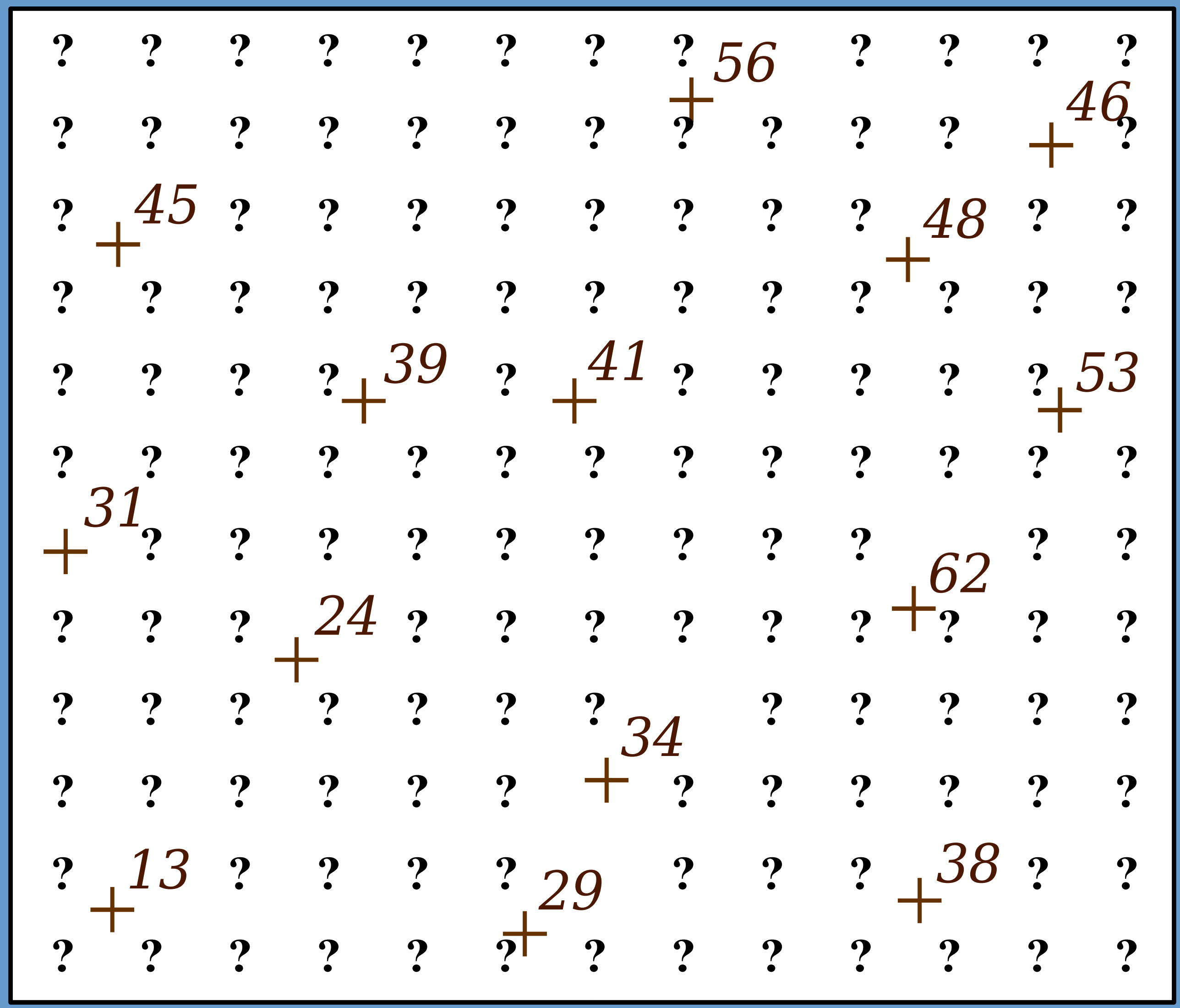

10
10
10
10
10
20
40
50


10
10
10
10
10
20
30
40
50

What is the missing value?

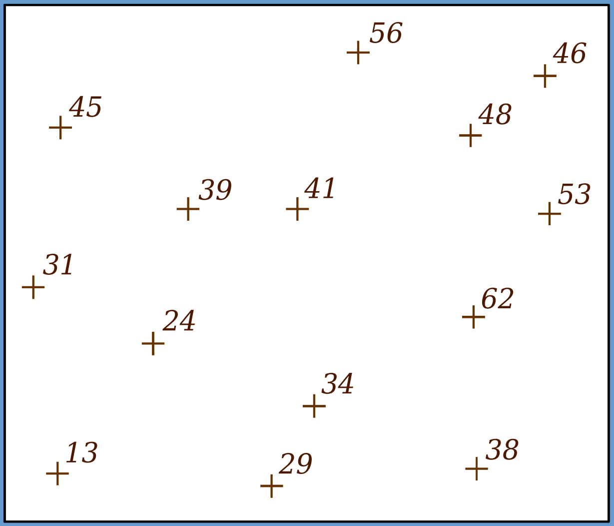

10
5
15
10
10
20
25
40
50

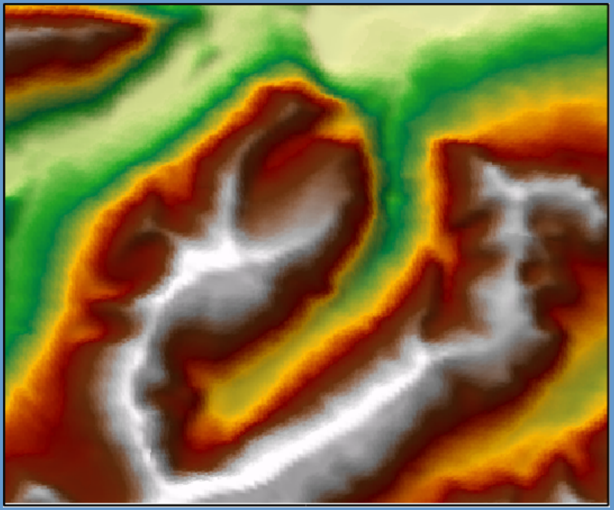
What are the missing values?



Interpolation

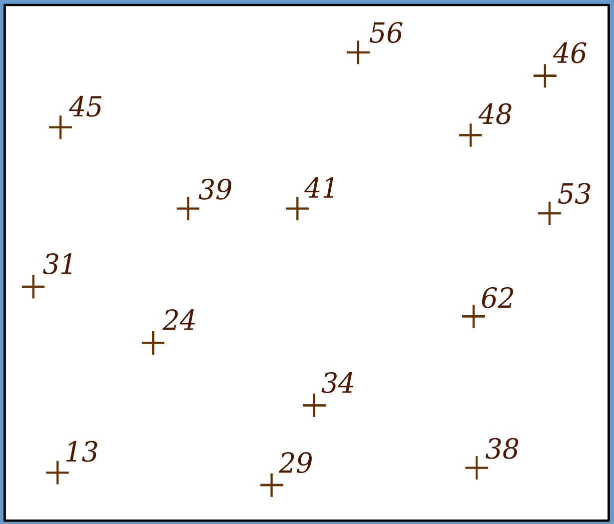
- *Interpolation* is the process of estimating attribute values at unmeasured locations
  - Especially useful when we only have *samples* distributed throughout space
  - Works for a single variable
    - Known values at specific locations
    - Interpolation "fills in the blanks"

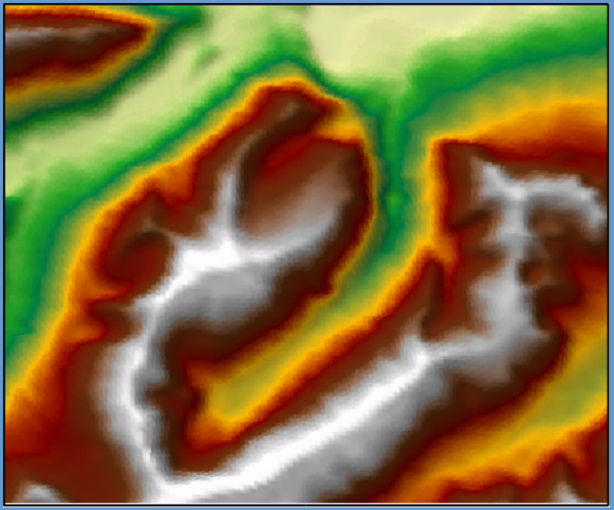




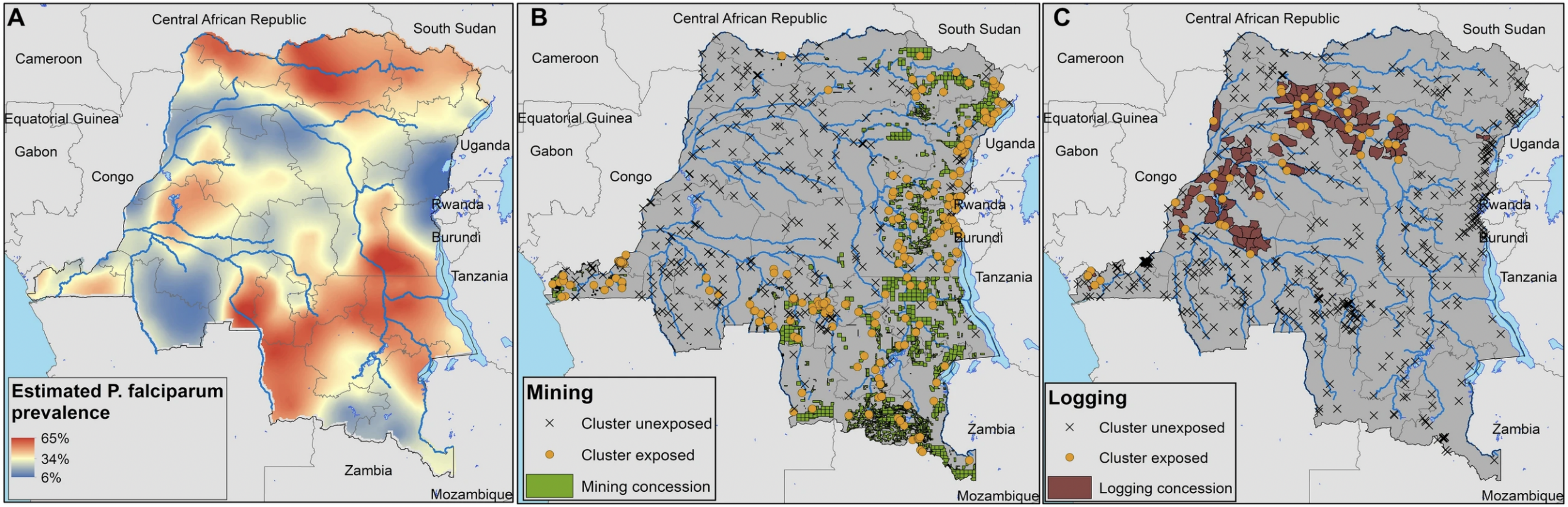
Interpolation

- *Interpolation* is the process of estimating attribute values at unmeasured locations
  - Takes advantage of our understanding of how values near to each other tend to be similar
    - Spatial autocorrelation





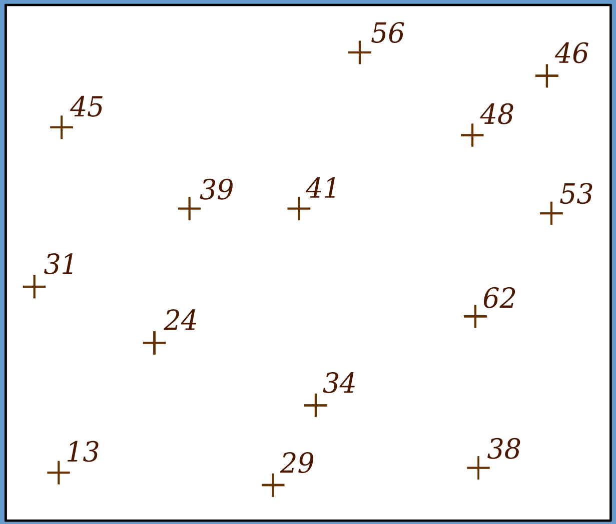
From: Impact of extractive industries on malaria prevalence in the Democratic Republic of the Congo: a population-based cross-sectional study

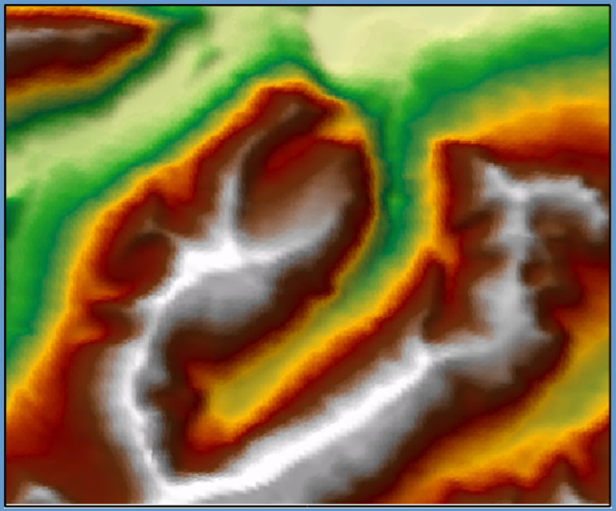


Mitchell et al., 2022

Interpolation

- *Interpolation* is the process of estimating attribute values at unmeasured locations
  - Assumes the underlying values are measured on a continuous scale
  - Assumes the underlying distribution is continuous across space



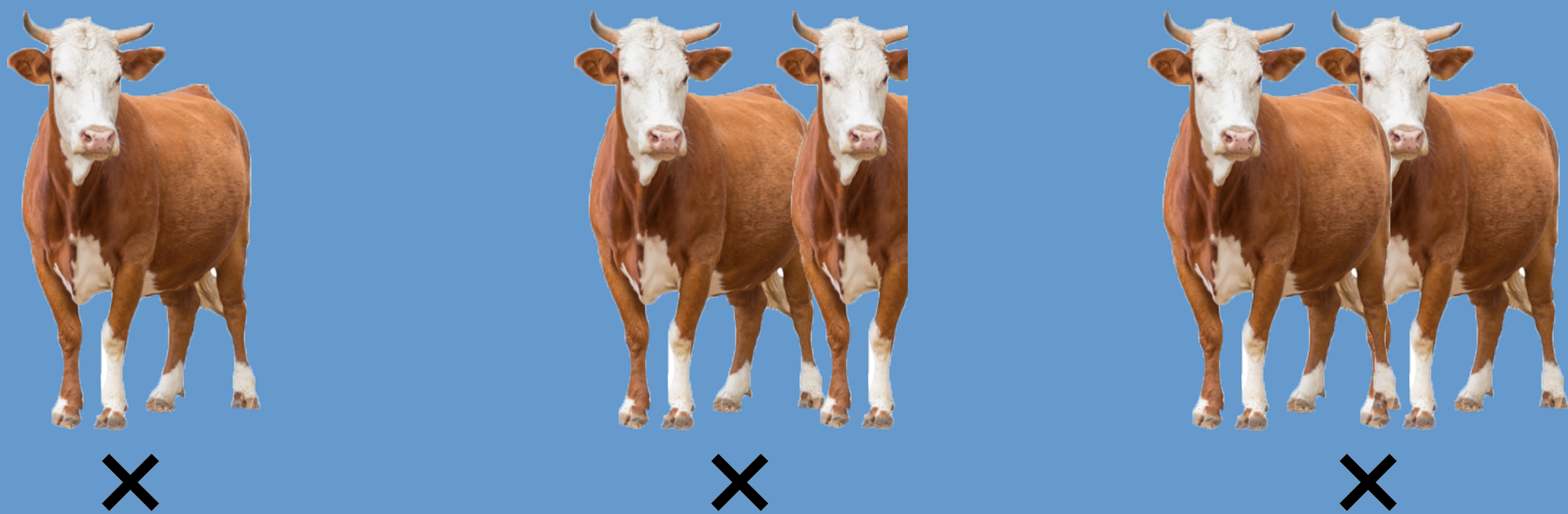


Cow Test!

Thanks to Dr. Tim Leslie

Cow Test!

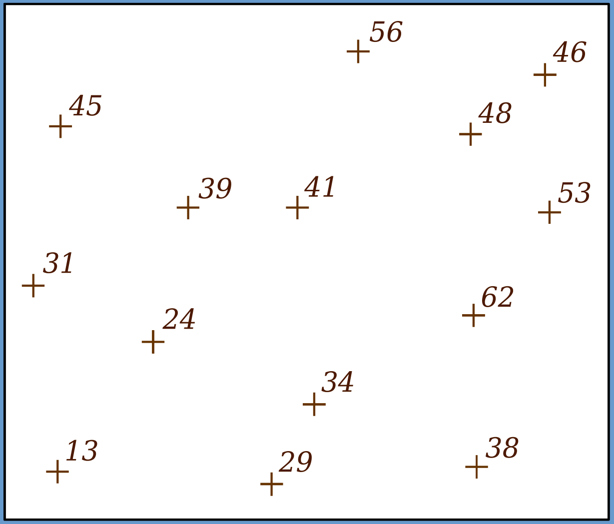
Thanks to Dr. Tim Leslie

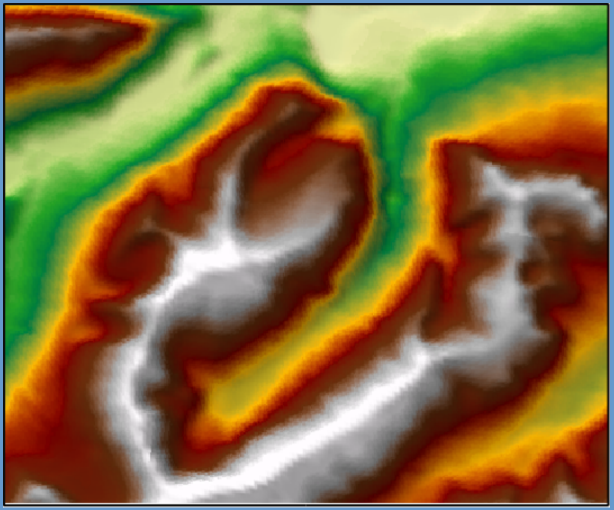


Does 1.5 cows make sense?

Interpolation

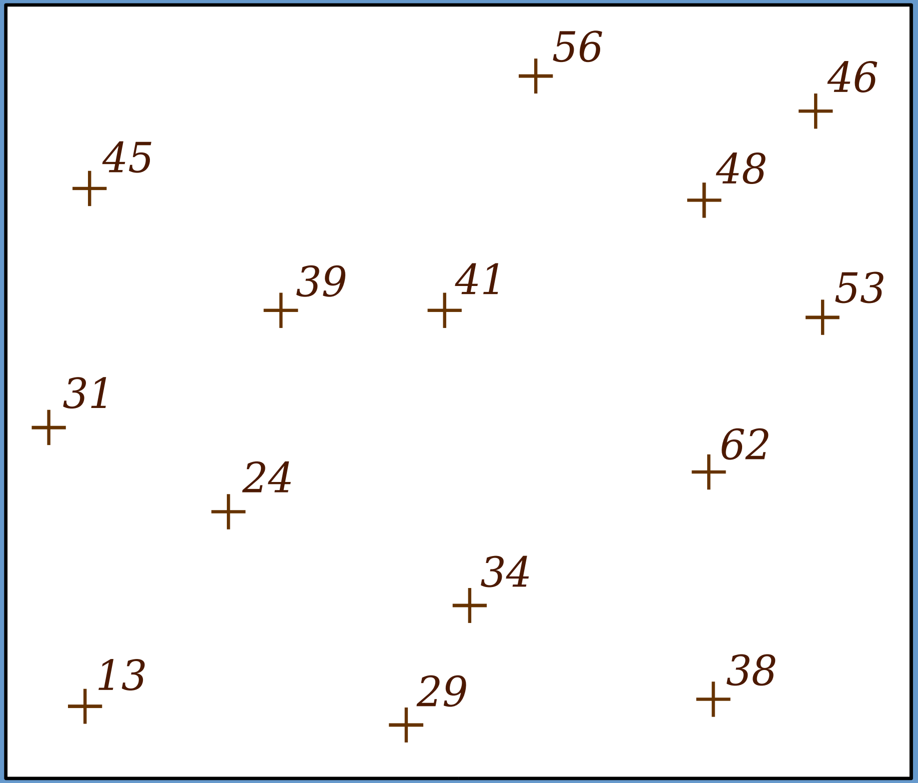
- *Interpolation* is the process of estimating attribute values at unmeasured locations
  - Useful for estimating the value of hazards at unknown locations
    - e.g., air pollution
  - Collection is only at discrete locations (points) but we want to estimate at non-sampled locations





Interpolation

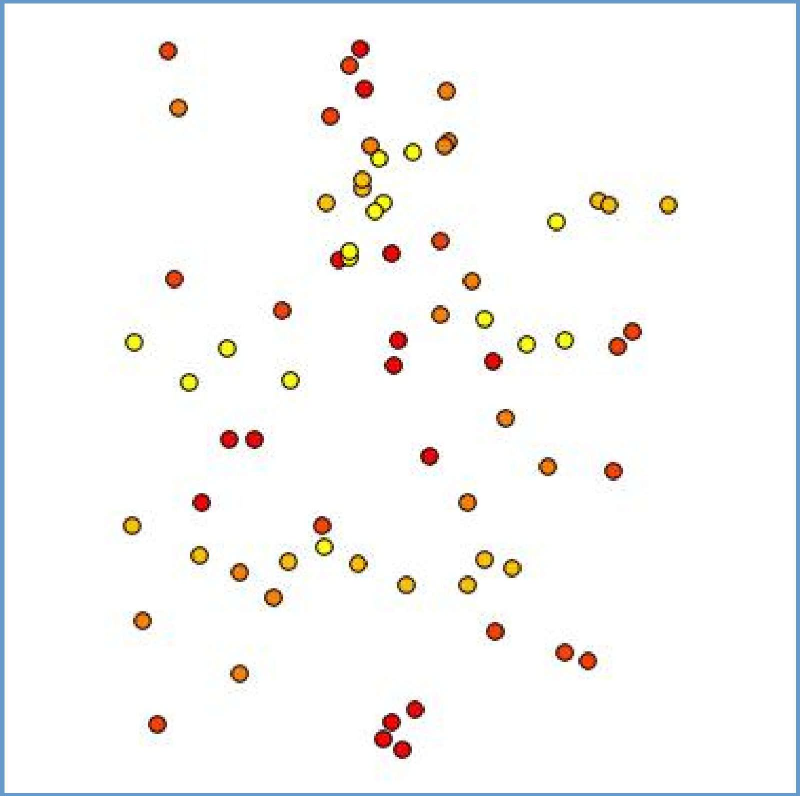
- Point-based approaches
  - Nearest neighbor
  - Trend surface
  - Inverse distance weighted (IDW)
  - Kriging



Input: set of point observations with measured attribute values
Output: continuous distribution of attribute values throughout a region

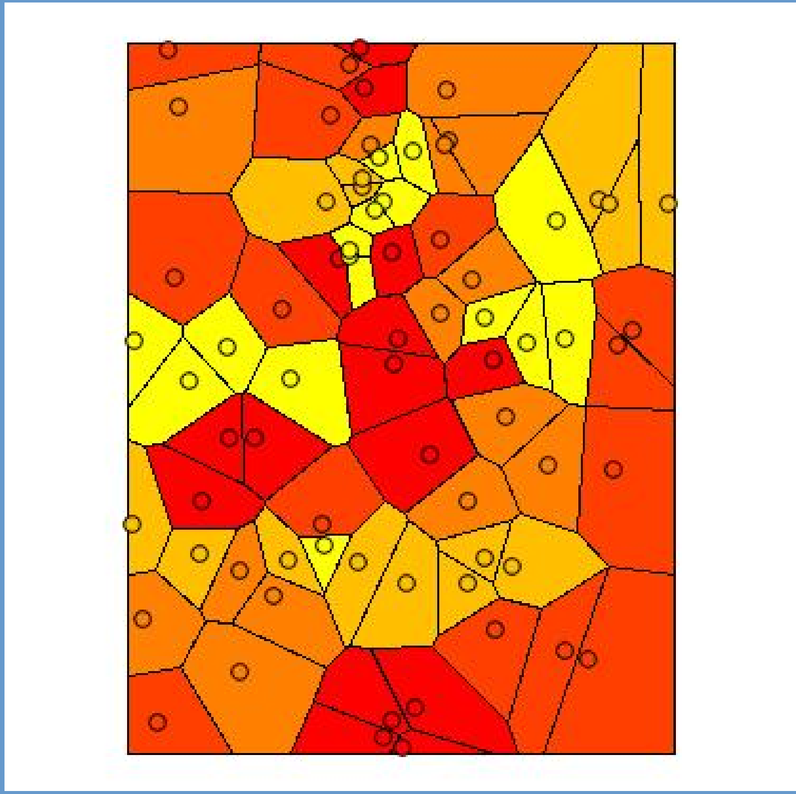
Nearest Neighbor

- Assigns each location to be interpolated the same value as the nearest sample point
  - Gives a stair-step shape to the surface (discontinuities)
- *Thiessen Polygons* are created when each location on the map is assigned to the nearest sample point



Nearest Neighbor

- Assigns each location to be interpolated the same value as the nearest sample point
  - Gives a stair-step shape to the surface (discontinuities)
- *Thiessen Polygons* are created when each location on the map is assigned to the nearest sample point

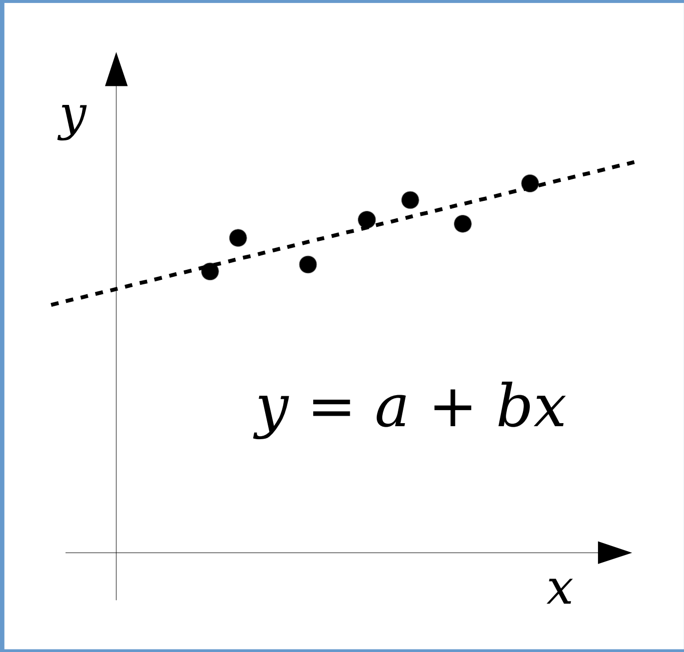


Nearest Neighbor

- Output features: polygons
- Estimated values are constrained to the input values
- At sample locations, estimated value will always equal sample value
- Results in a “shelved” surface

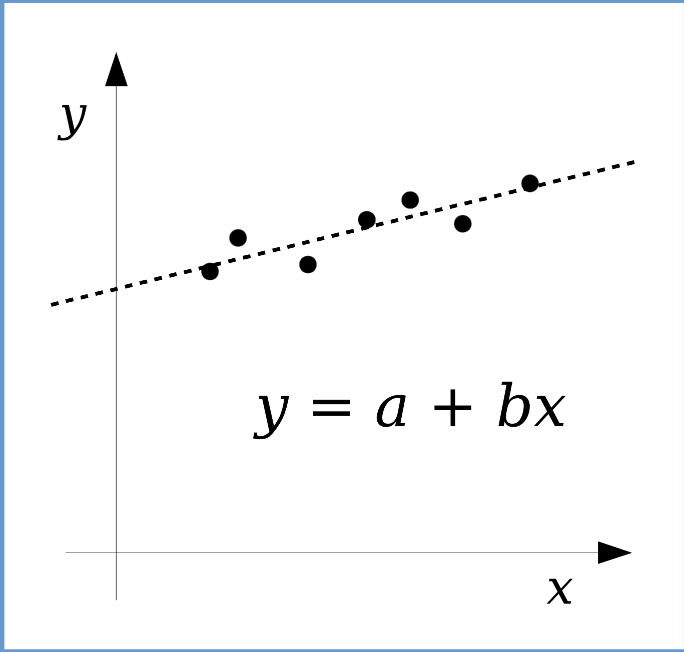
Trend Surface Analysis

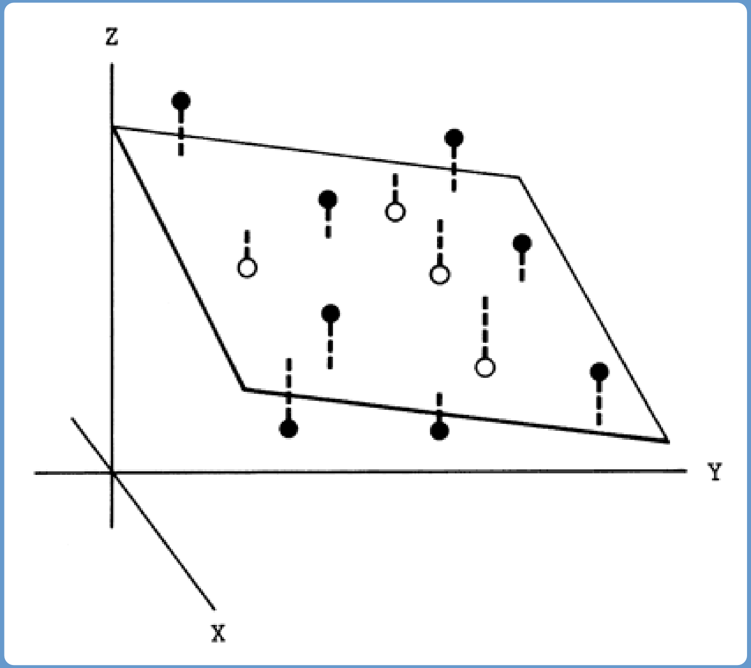
- Fit a plane or other surface through the observed points
  - Similar to fitting a regression line to a set of observations



Trend Surface Analysis

- Fit a plane or other surface through the observed points
  - Similar to fitting a regression line to a set of observations



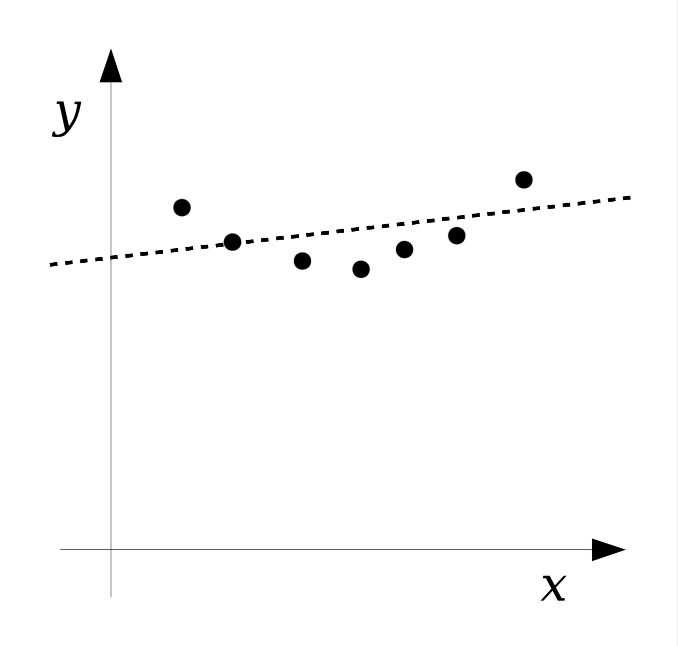


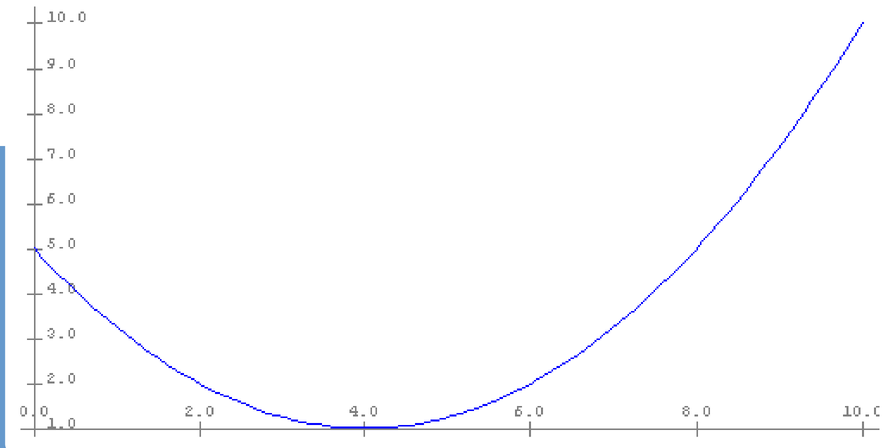
- First order polynomial equation

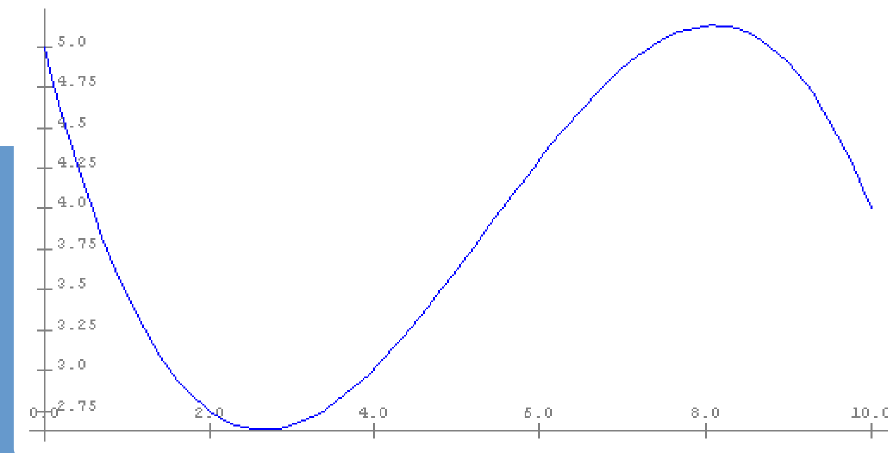
$$z(x,y) = a + bx + cy$$

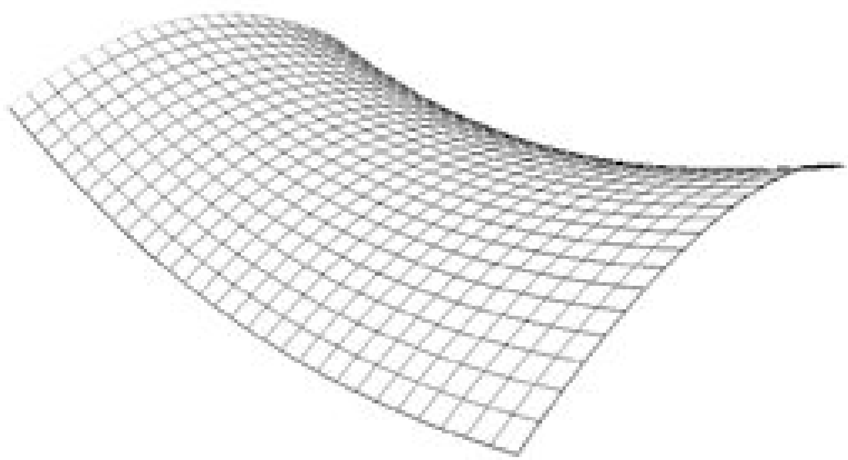
Trend Surface Analysis

- Fit a plane or other surface through the observed points
  - Similar to fitting a regression line to a set of observations

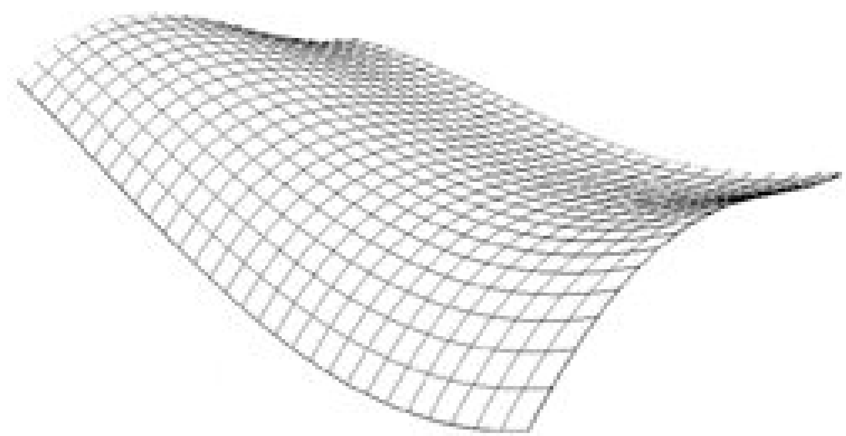








d) 2nd order trend surface



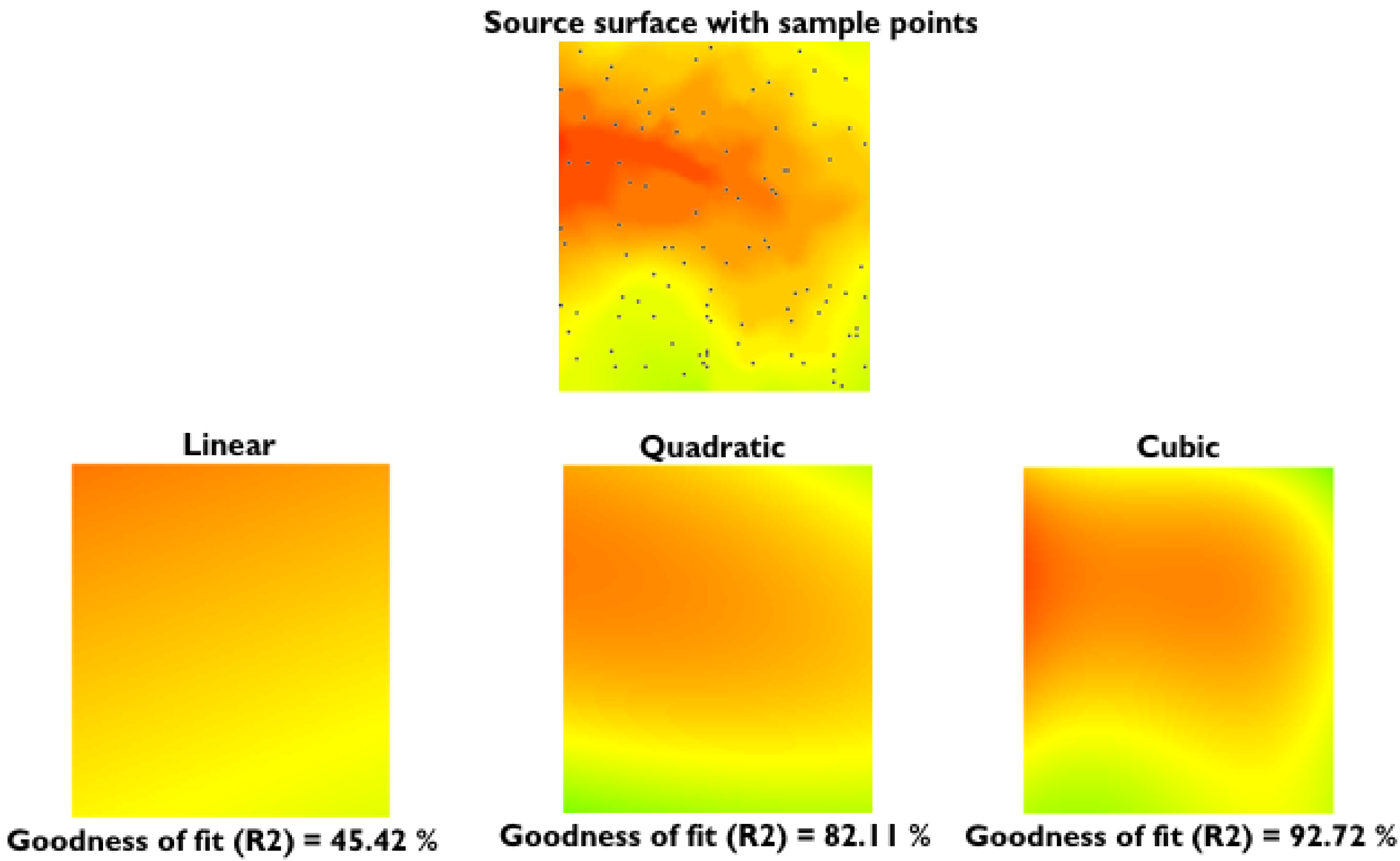
e) 3rd order trend surface

- Second order polynomial equation

$$z(x,y) = a + bx + cy + dxy + ex^2 + fy^2$$

Trend Surface Analysis

- Fitting a trend surface to observations
- Order of the equation and sample points
  - Limit on the number of coefficients: you cannot have more than the number of sample points
    - Just because you can doesn't mean you should!
  - Surface fit and model residuals
    - Regression diagnostics
      - R^2, *f*, residual plot



www.geog.leeds.ac.uk/courses/level2/geog2750/geog2750_14.ppt

Trend Surface Analysis

- Output features: raster (in reality, a function)
- Estimated values are not constrained to the input values
- At sample locations, estimated value does not always equal sample value
- Results in a very smooth surface

IDW

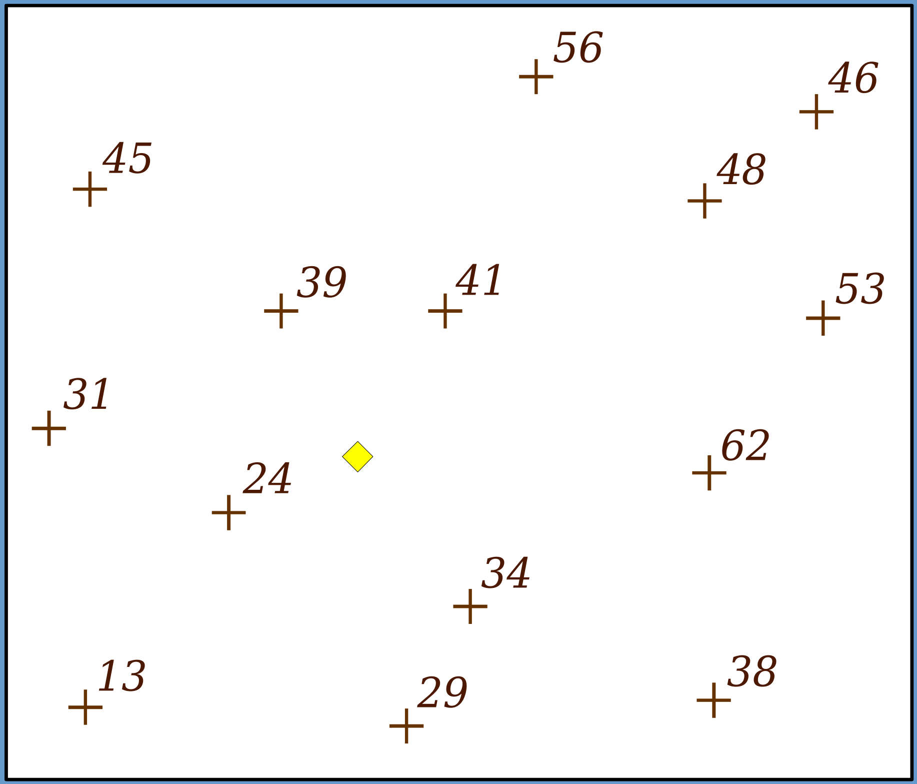
- Inverse Distance Weighted interpolation
  - Value for non-sampled locations is calculated as the weighted average of nearby values
  - Weighted average often takes the form:

$$z_j = \frac{\sum_{i=1}^{n} \frac{z_i}{d_i^2}}{\sum_{i=1}^{n} \frac{1}{d_i^2}}$$

  - *Requires*: neighborhood to be defined (n)
    - Specific number of points or specified *d* (radius)
  - *Key assumption*: the value at a point is assumed to be related to the value at other points such that the similarity declines at a rate of $1/d^x$
    - $1/d^x$ is only a guess at how nearby values are related ($1/d^2$ is commonly used)

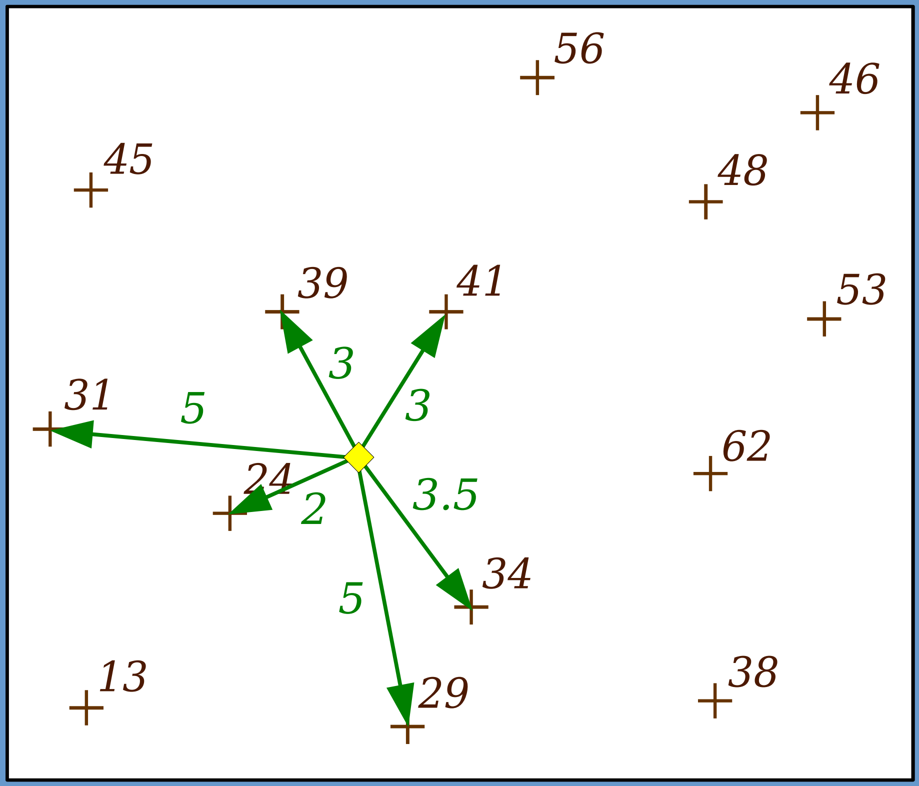
IDW Calculation

- Calculate value at the yellow diamond
  - $n = 6$



IDW Calculation

- Calculate value at the yellow diamond
  - $n = 6$



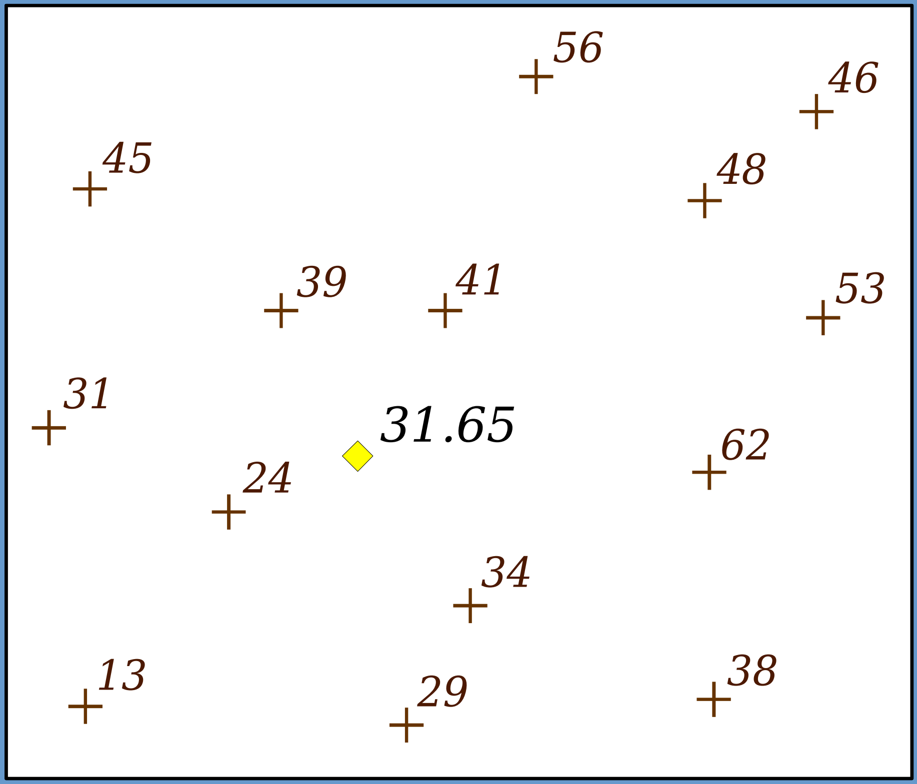
$$z_j = \frac{\frac{24}{2^2} + \frac{39}{3^2} + \frac{41}{3^2} + \frac{34}{3.5^2} + \frac{31}{5^2} + \frac{29}{5^2}}{\frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{3^2} + \frac{1}{3.5^2} + \frac{1}{5^2} + \frac{1}{5^2}}$$

$$z_j = \frac{20.0644}{0.6339}$$

$$z_j = 31.6546$$

IDW Calculation

- Calculate value at the yellow diamond
  - $n = 6$



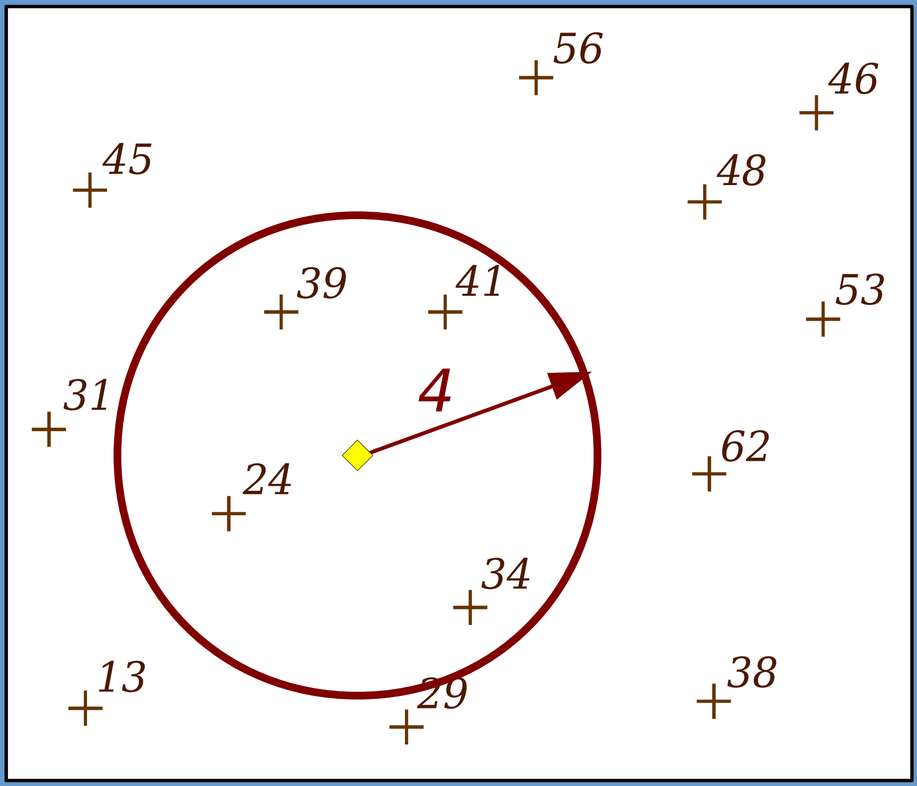
$$z_j = \frac{\frac{24}{2^2} + \frac{39}{3^2} + \frac{41}{3^2} + \frac{34}{3.5^2} + \frac{31}{5^2} + \frac{29}{5^2}}{\frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{3^2} + \frac{1}{3.5^2} + \frac{1}{5^2} + \frac{1}{5^2}}$$

$$z_j = \frac{20.0644}{0.6339}$$

$$z_j = 31.6546$$

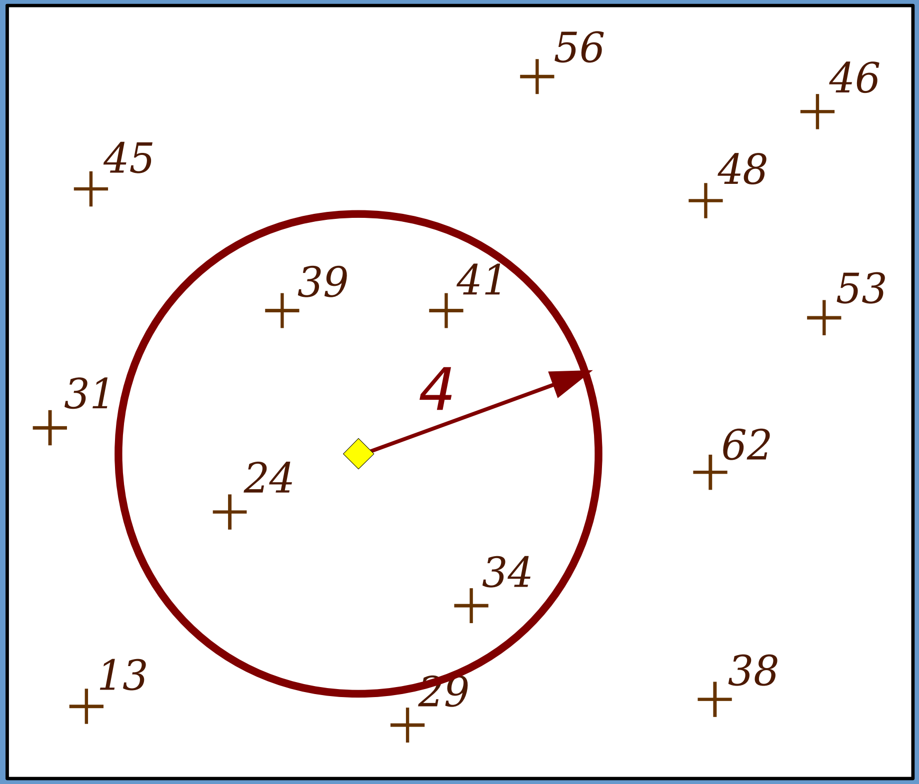
IDW Calculation

- Calculate value at the yellow diamond
  - $d = 4$



IDW Calculation

- Calculate value at the yellow diamond
  - $d = 4$



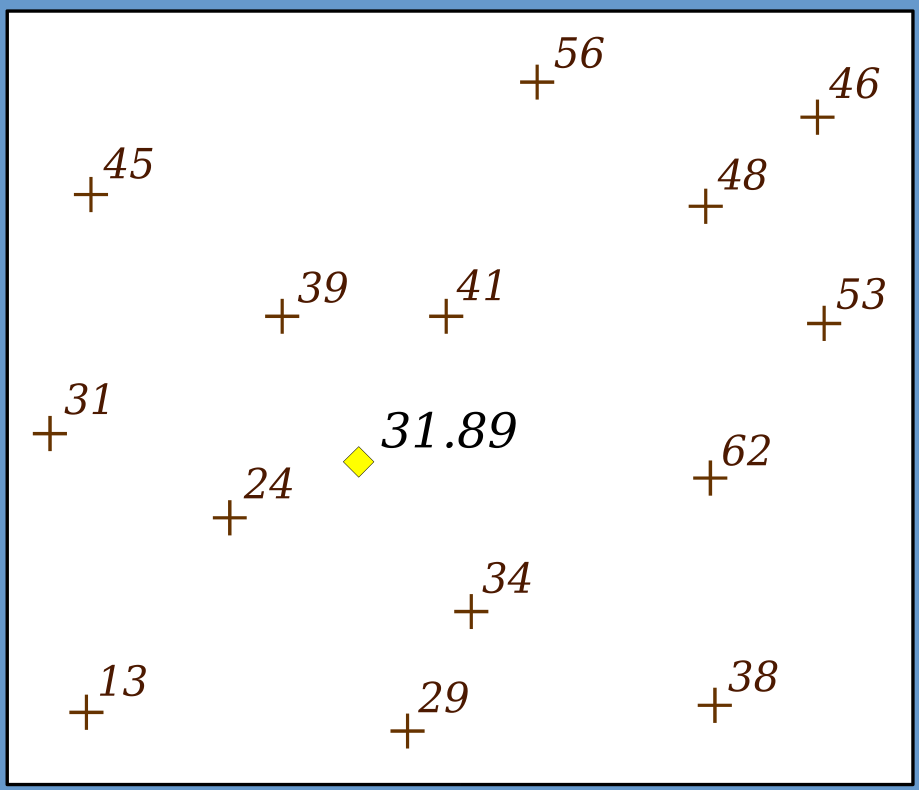
$$z_j = \frac{\frac{24}{2^2} + \frac{39}{3^2} + \frac{41}{3^2} + \frac{34}{3.5^2}}{\frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{3^2} + \frac{1}{3.5^2}}$$

$$z_j = \frac{17.6644}{0.5539}$$

$$z_j = 31.8936$$

IDW Calculation

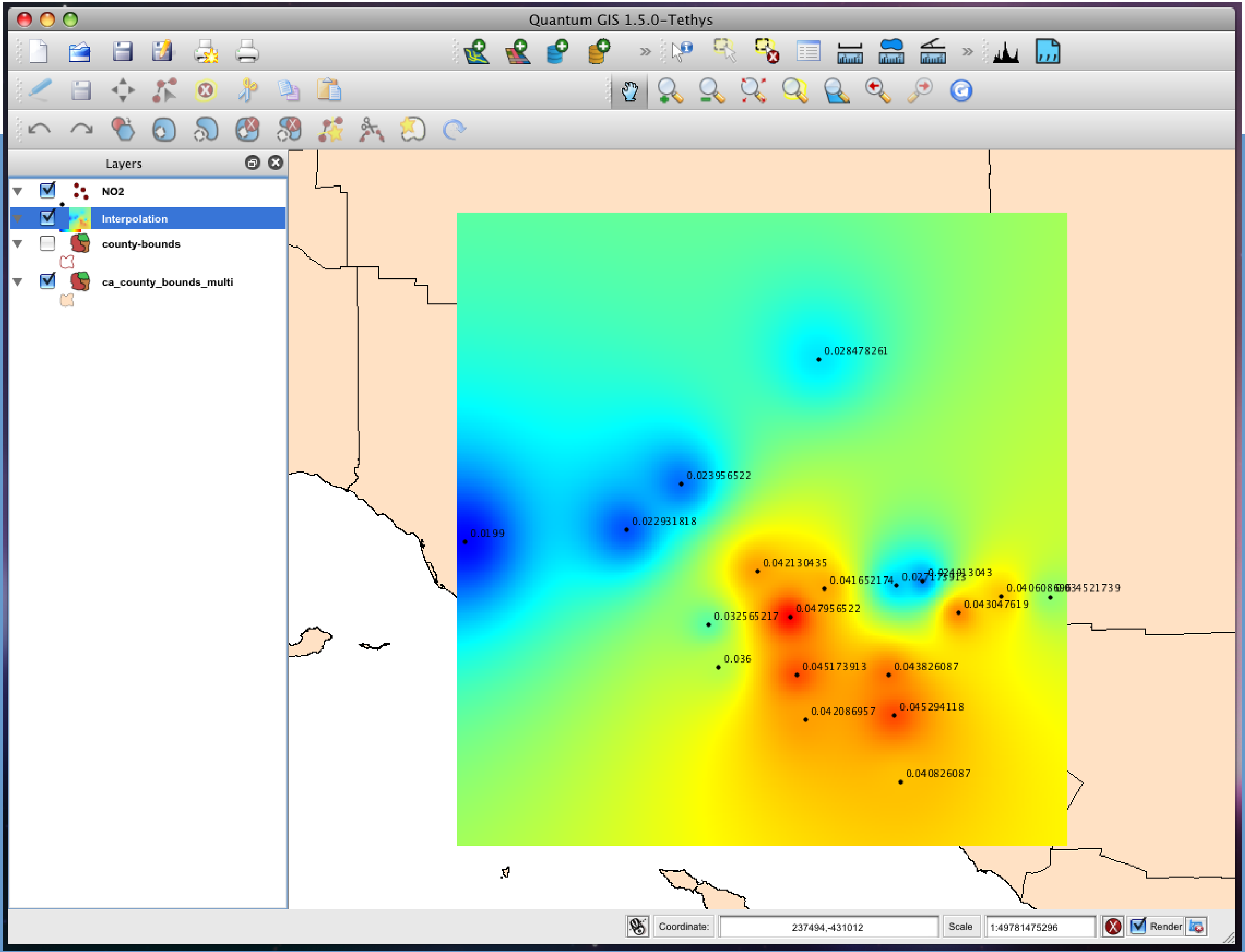
- Calculate value at the yellow diamond
  - $d = 4$



$$z_j = \frac{\frac{24}{2^2} + \frac{39}{3^2} + \frac{41}{3^2} + \frac{34}{3.5^2}}{\frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{3^2} + \frac{1}{3.5^2}}$$

$$z_j = \frac{17.6644}{0.5539}$$

$$z_j = 31.8936$$


Quantum GIS 1.5.0-Tethys
Layers
NO2
Interpolation
county-bounds
ca_county_bounds_multi
0.028478261
0.023956522
0.022931818
0.0199
0.042130435
0.041652174
0.047956522
0.032565217
0.036
0.045173913
0.043826087
0.042086957
0.045294118
0.040826087
0.043047619
Coordinate:
237494,-431012
Scale
1:49781475296
Render

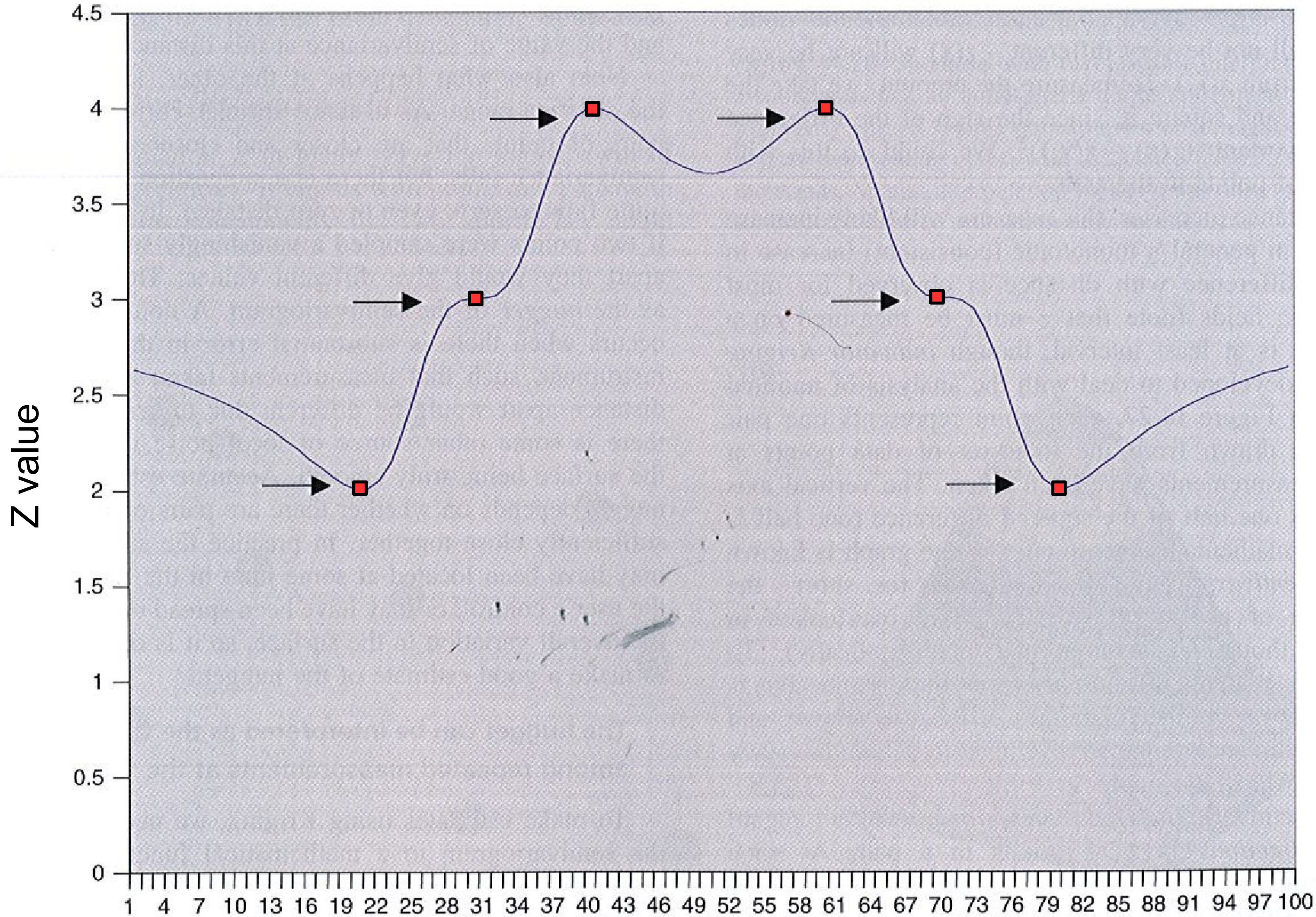


Figure 14.26 Potentially undesirable characteristics of IDW interpolation. Data points located at 20, 30, 40, 60, 70, and 80 have measured values of 2, 3, 4, 4, 3, and 2 respectively. The interpolated profile shows a pit between the two highest values, and regression to the overall mean value of 3 outside the area covered by the data

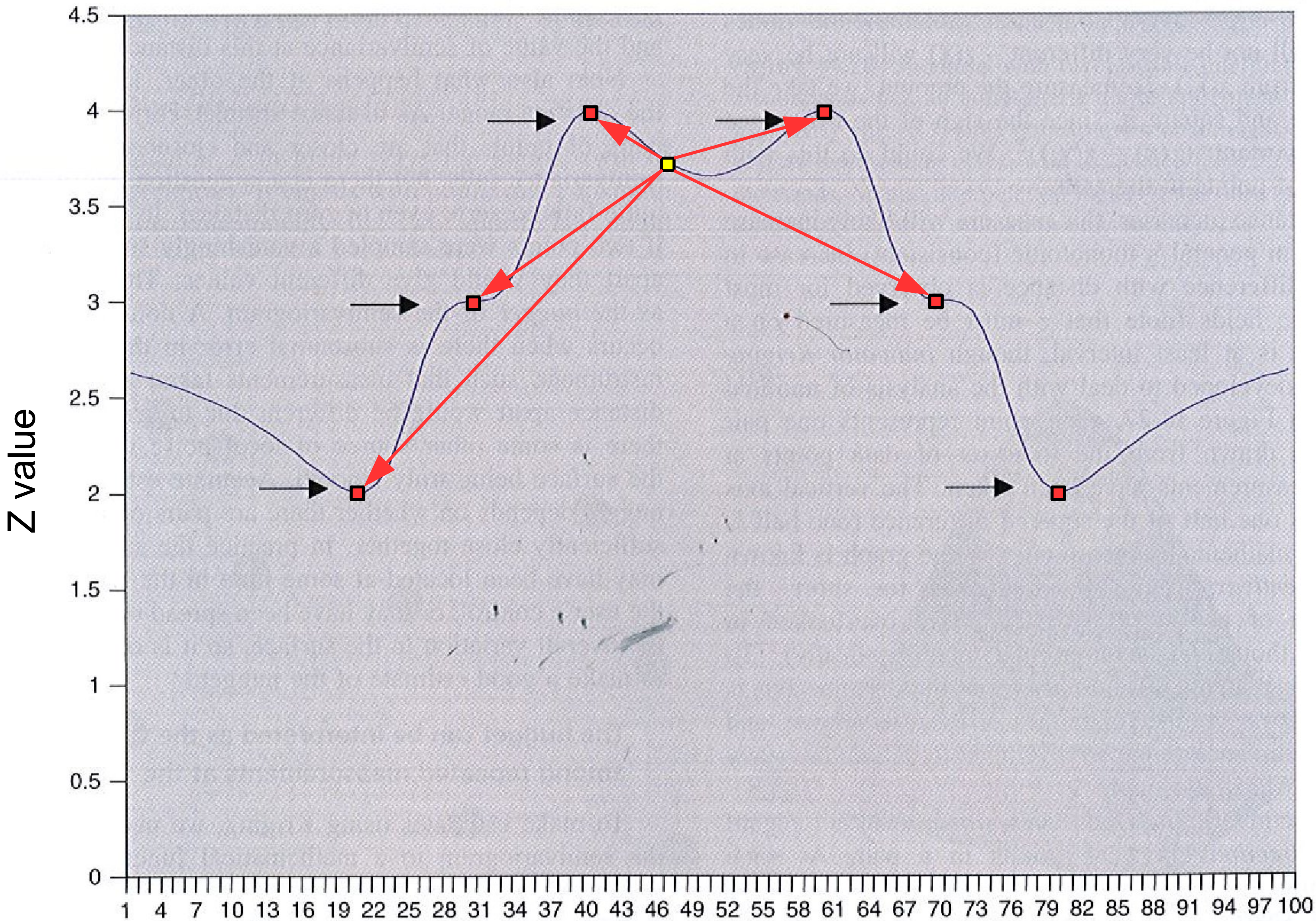
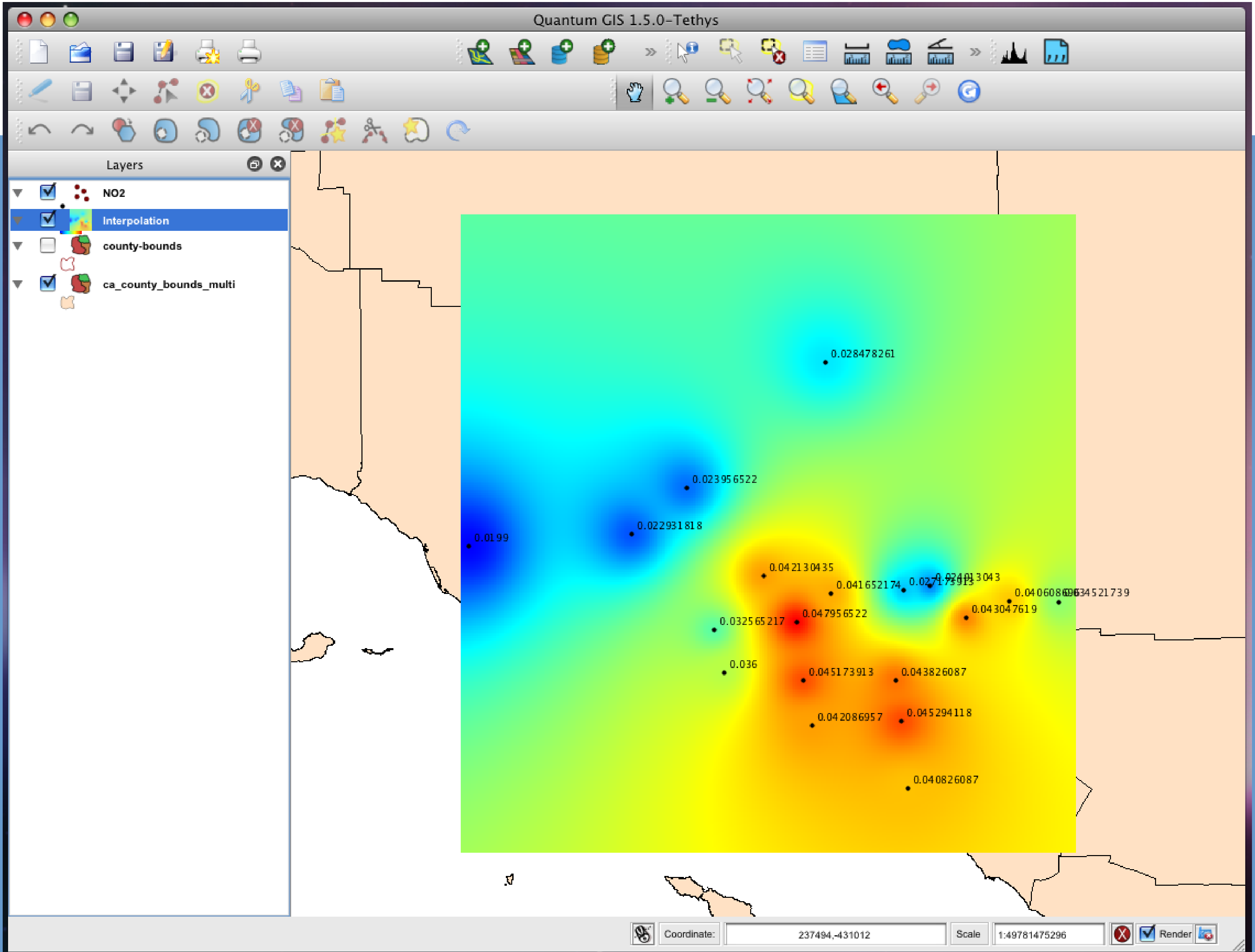


Figure 14.26 Potentially undesirable characteristics of IDW interpolation. Data points located at 20, 30, 40, 60, 70, and 80 have measured values of 2, 3, 4, 4, 3, and 2 respectively. The interpolated profile shows a pit between the two highest values, and regression to the overall mean value of 3 outside the area covered by the data



Quantum GIS 1.5.0-Tethys
Layers
NO2
Interpolation
county-bounds
ca_county_bounds_multi
0.028478261
0.023956522
0.022931818
0.0199
0.042130435
0.041652174
0.032565217
0.047956522
0.043047619
0.036
0.045173913
0.043826087
0.042086957
0.045294118
0.040826087
Coordinate:
237494,-431012
Scale
1:49781475296
Render

IDW

- Output features: raster
- Estimated values are constrained to the range of the input values
- At sample locations, estimated value always equals sample value
- Results in a pitted surface (local highs and lows)

Kriging

- Uses a *distance weighted average* method to calculate the value at unsampled locations
  - Weights are estimated from the data using the *semivariogram*
    - *Rather than guessing at the relationship* between similarity of values and distance (like we do when we guess at the exponent in IDW)

Kriging

- Called a *geostatistical* method
  - Parameters based on input data
    - Observed similarity as a function of distance
  - Statistical, not mathematical
    - Provide a measure of uncertainty

Semivariance

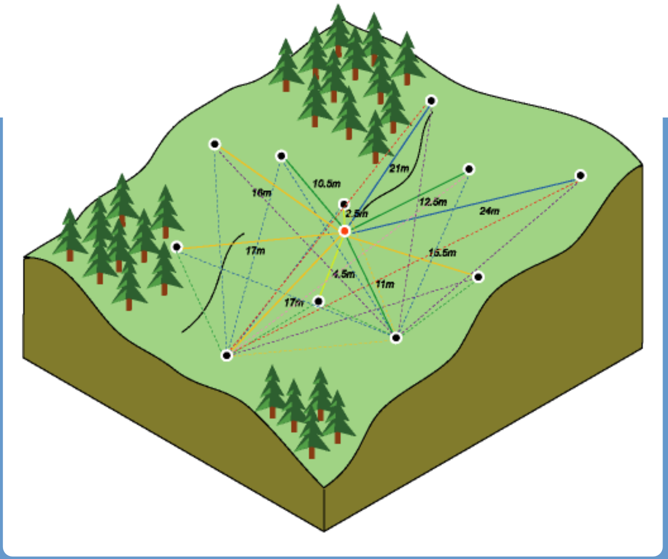
- *Semivariance*: the observed average variation between values measured at some separation distance (h)

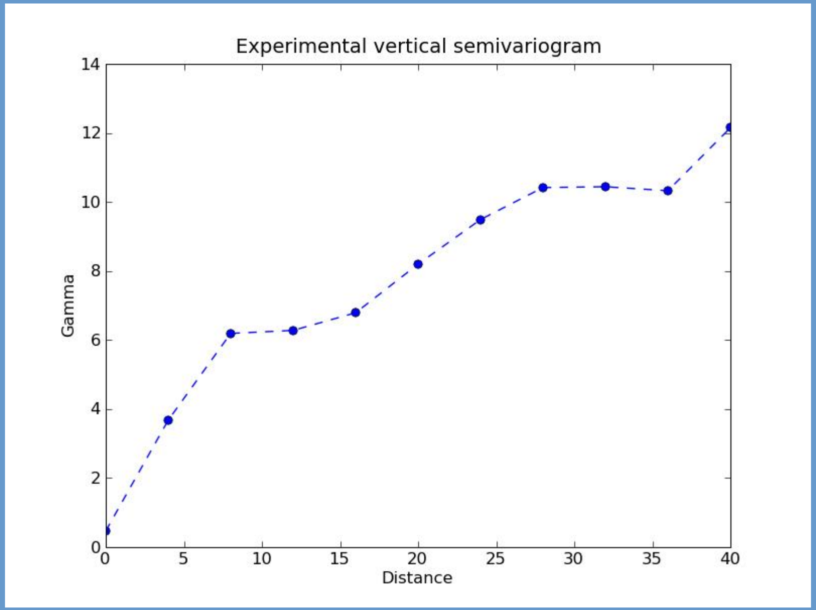
$$\gamma(h) = \frac{1}{2n} \sum_{i=1}^{n} \left[z(x_i) - z(x_i + h) \right]^2$$

- *Semivariogram*: the observed relationship between semivariance and distance (h)
  - Provides a description of spatial structure
  - Called "omnidirectional" when calculated regardless of any directional effects

Semivariogram

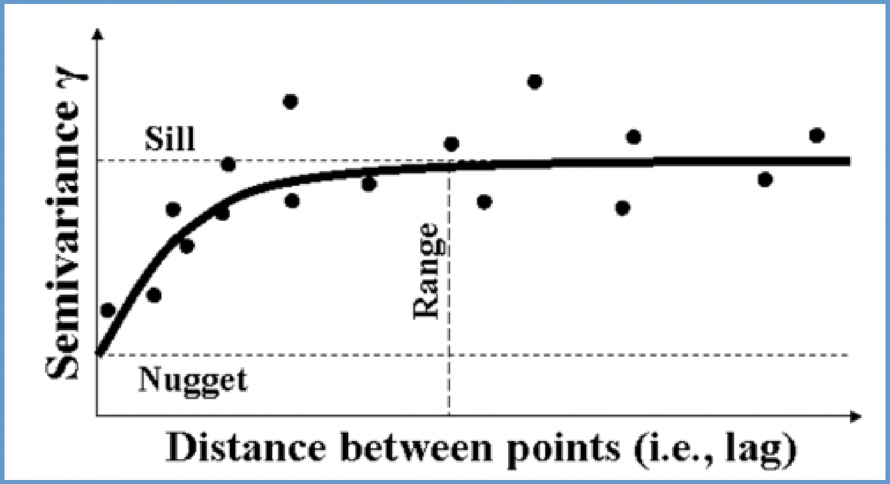
- In order to know how similarity changes with distance we calculate the empirical semivariogram from the sample data
  - Its shape tells us something about the general form of the surface (or continuous field)
  - Typically shows that difference between values increases with distance

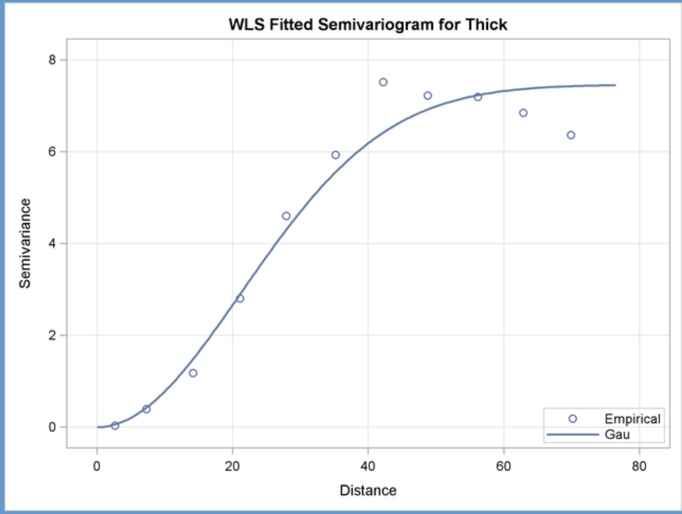




Semivariogram

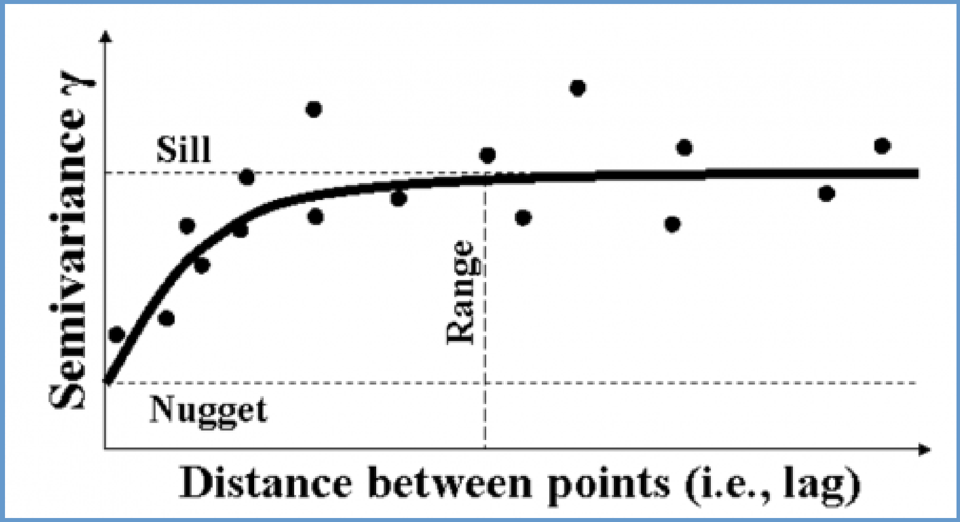
- Modeling the semivariogram
  - Describe the relationship between difference of values (variation) and distance using a mathematical function
  - Several different types of functions can be used, each with a different form to the distance function





Modeling the Semivariogram

- Described quantitatively by sill, range, and nugget and using different function forms

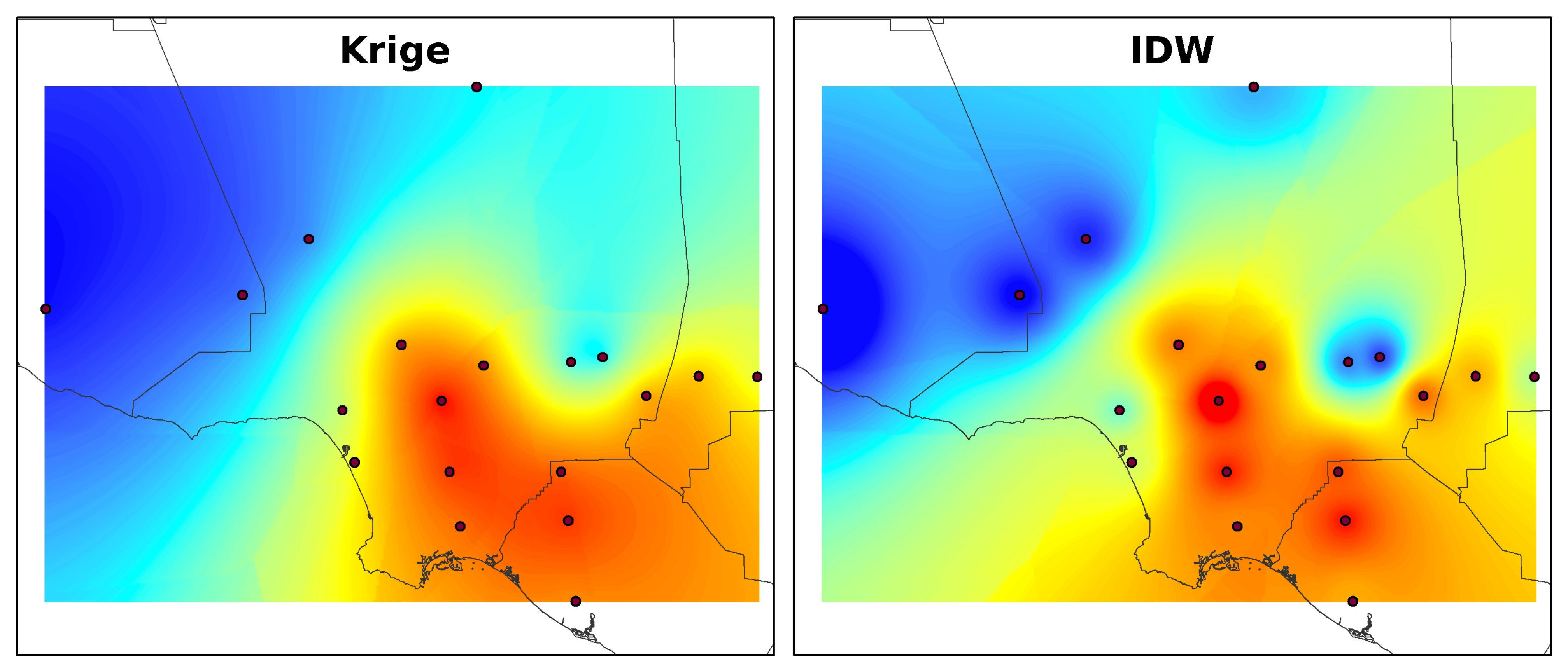


- *Range*: distance beyond which values are not spatially autocorrelated
- *Sill*: the true variance of the variable, observed between variables separated by a distance > range
- *Nugget*: the level of variance observed when the semivariogram curve is projected back to zero distance

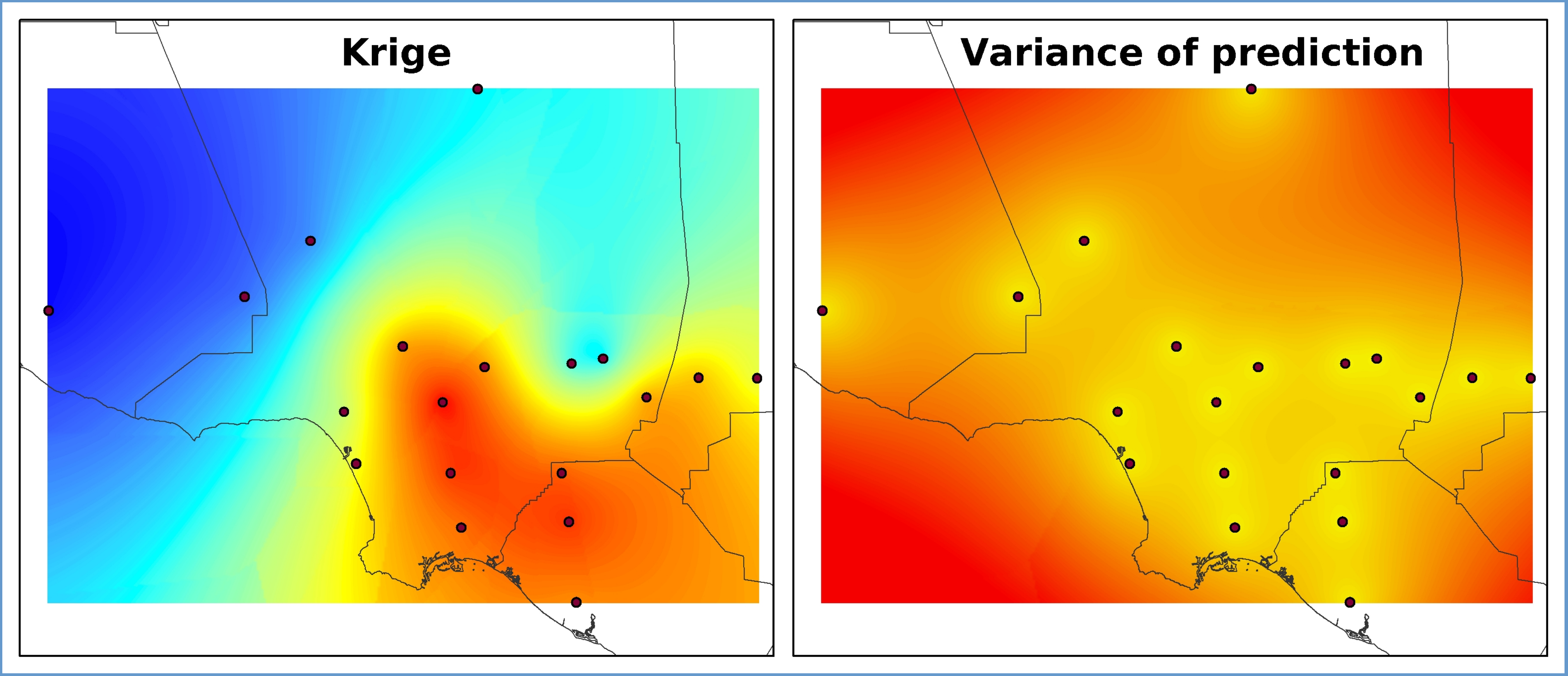
Kriging

- Estimates the weights based on distance
  - Using mathematical function derived from the semivariogram
- Calculates the weighted average of neighboring values for each location
  - Neighbors values as modified by distance and weight (associated with distance)
  - Kriging can be very computationally intensive
- *Provides interpolated surface plus an surface of uncertainty of prediction*

Kriging



Kriging



Kriging

- Output features: raster (estimated value and uncertainty)
- Estimated values are not constrained to the range of the input values
- At sample locations, estimated value not always equal to sample value
- Results in a smoother surface (than IDW)

Non-stationarity

- Thus, far the methods assume that there is a global spatial autocorrelation structure in the data
  - e.g., 1 / *d* applied to all places in IDW
  - e.g., global semivariogram function in Kriging
- This may not be appropriate for variables/phenomena that have a highly variable spatial structure
  - e.g., Think about how temperature varies in plains vs mountains (orographic effects)

Empirical Bayesian Kriging

- Does not assume a stationary process
- Accounts for error introduced by estimating a global semivariogram model
  - Estimates numerous local semivariogram models
    - Captures variation in the spatial dependence from place to place in the study region

Empirical Bayesian Kriging

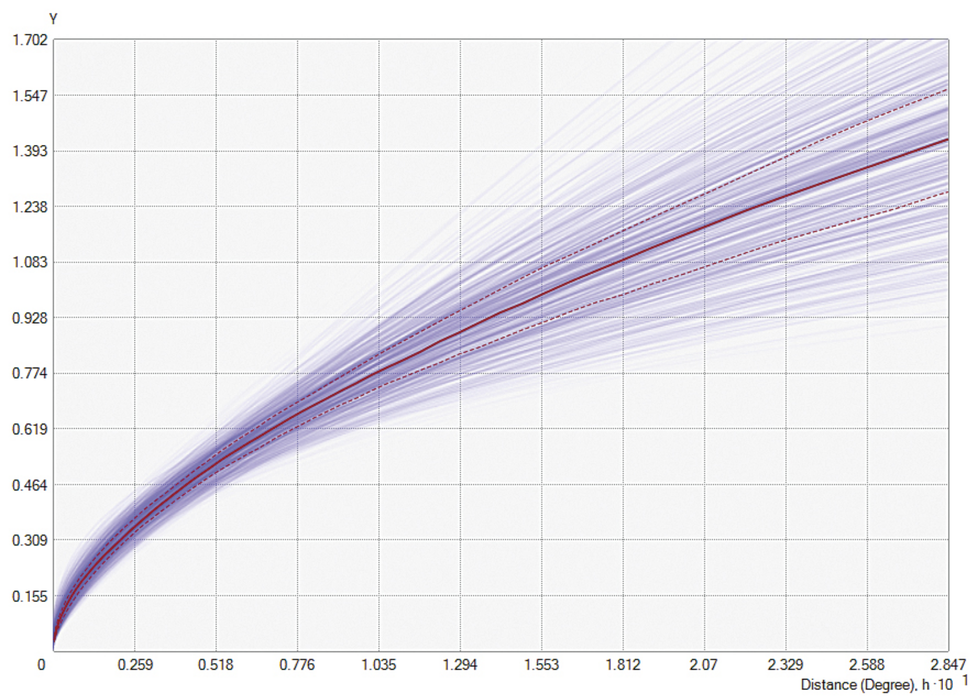
- Subset observed data to local region
  - Estimate initial semivariogram model from observed data values
  - Use initial semivariogram model to simulate a value at each original data point
  - Estimate new semivariogram model from simulated data values
    - Calculate the likelihood that the original data values were generated from the new semivariogram and assign weight (Bayes' rule)

Repeated Simulations

Empirical Bayesian Kriging

- Creates a large number of semivariograms for each subset (local region)
  - When plotted together, the result is a distribution of semivariograms

Example of distribution of Semivariograms: The darker the blue color, the more semivariogram models are present



Empirical Bayesian Kriging

- Prediction: for each unmeasured location, calculate a new semivariogram distribution using likelihood-based sampling of individual semivariograms
  - Using weight values
  - Can be drawn from multiple spatial subsets
    - For example, if a prediction location has neighbors with observed values in two different subsets, the prediction will be calculated using simulated semivariograms from both of the subsets
    - Semivariograms are chosen based on their likelihood values

EB Kriging

- Output features: raster (estimated value and uncertainty)
- Estimated values are not constrained to the range of the input values
- At sample locations, estimated value not always equal to sample value
- Results in a different surface (than classical kriging), depending on whether non-stationary

Keywords

- Interpolation
- Samples
- Unmeasured locations
- Cow test
- Point-baed
- Thiessen polygons
- Trend surface
- IDW
- Kriging
- Semivariogram
- Geostatistical
- Uncertainty