Lecture 8: Spatial Interpolation (IDW)

GGR376

Dr. Adams

What is it?

A method of constructing new data points within the range of a discrete set of known data points.

Example - Maple Sap Collection

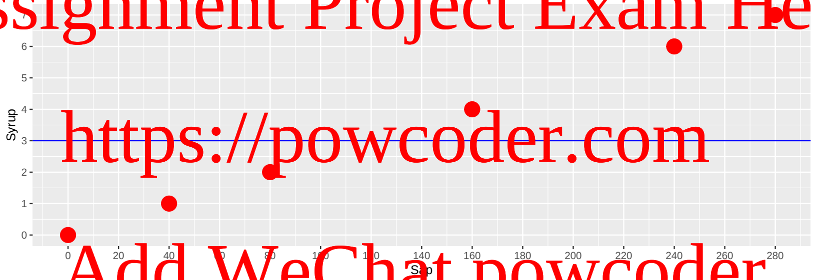
Example - Daily Sap Collection

Create a plot of Day vs Sap Collected

Day	Total Sap Collected (Gallons)
0	0
1	4
2	NA
3	12
4	16
6	20
7	NA
8	28

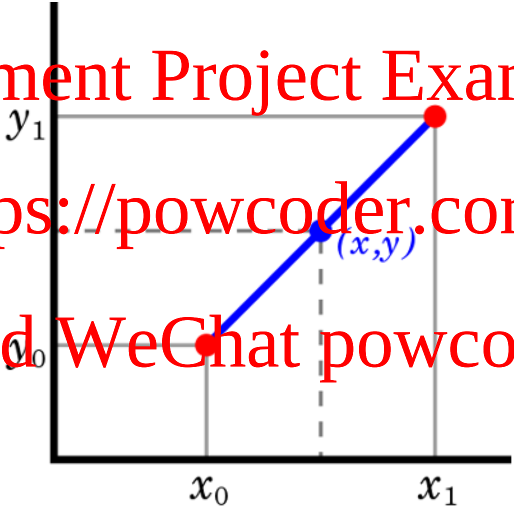
How much sap should have been collected by Day 2 and Day 7?

Example - Sap Collected Syrup Produced



How much sap is required for 3 litres of syrup?

Linear Interpolation



Linear Interpolation Equation

Given two known coordinates:

(x_0, y_0) and (x_1, y_1)

$$y = \frac{y_0(x_1 - x) + y_1(x - x_0)}{x_1 - x_0}$$

Try

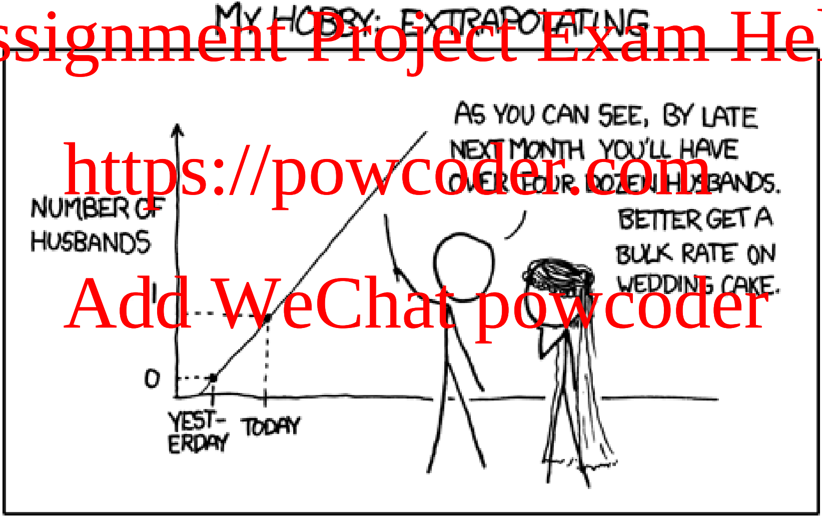
Given (3, 5) & (5, 10) What is $x = 4$

Liner Interpolation Recap

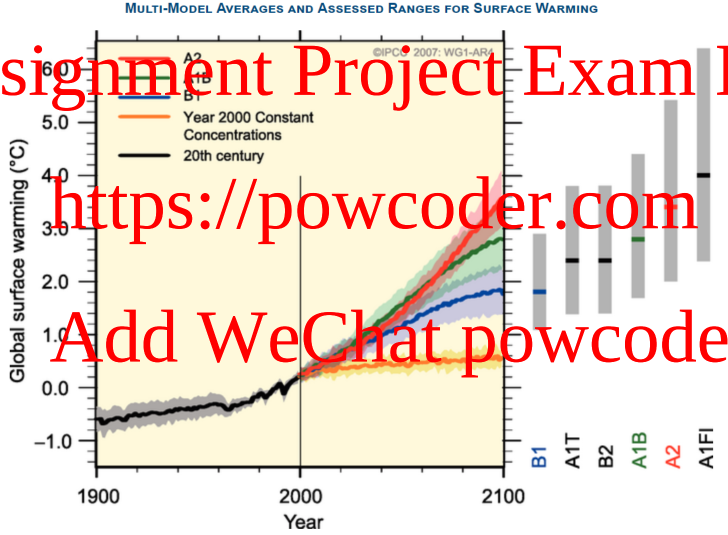
- One dimension
- Linear approach
- Within the range of the data

Outside the range

Extrapolation: Estimating beyond the original observation range.



IPCC: Projections of Future Changes in Climate



What about space?

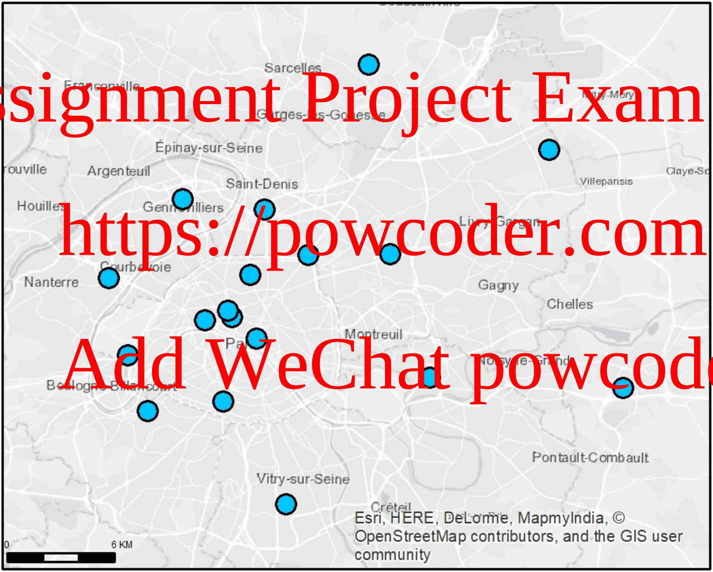
"Everything is related to everything else, but near things are more related than distant things." - Waldo Tobler

Spatial Interpolation

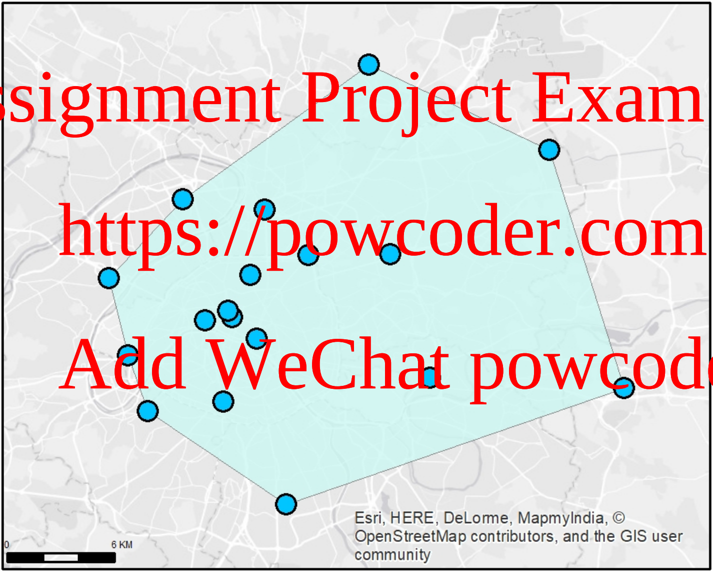
Prediction of values at unknown locations

- Rainfall
- Air Pollution
- Ground Water Depth
- Elevation

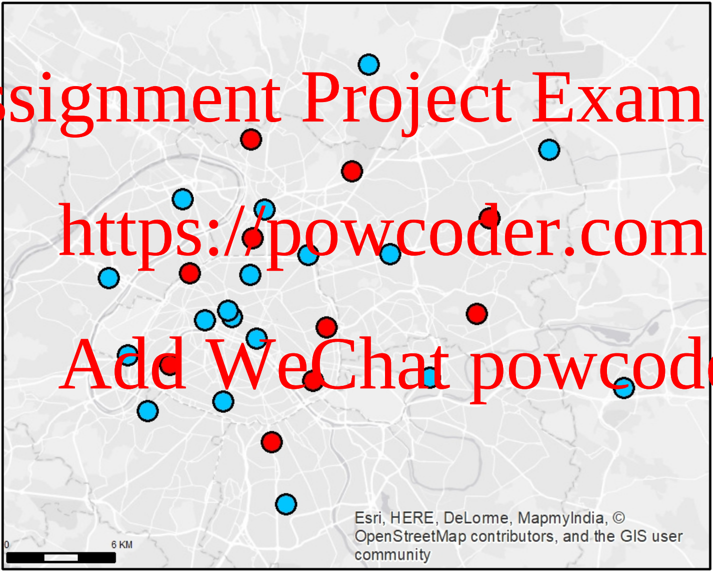
Paris, France - Air Pollution Monitors



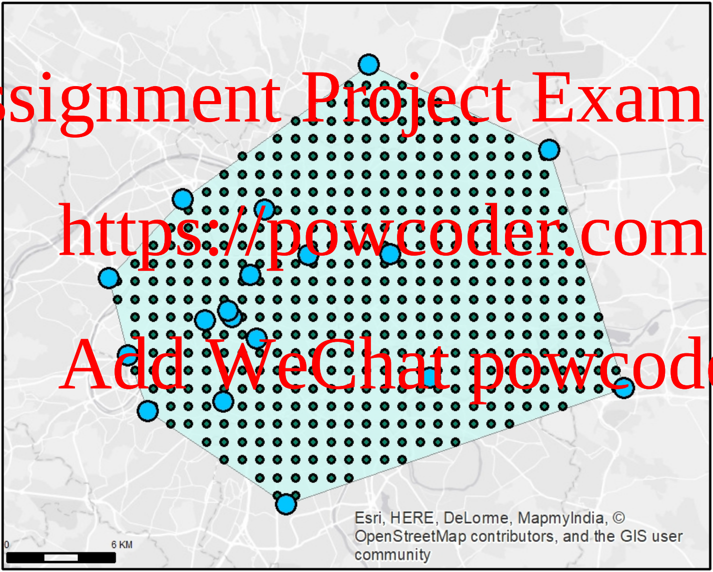
Paris - Interpolation vs. Extrapolation



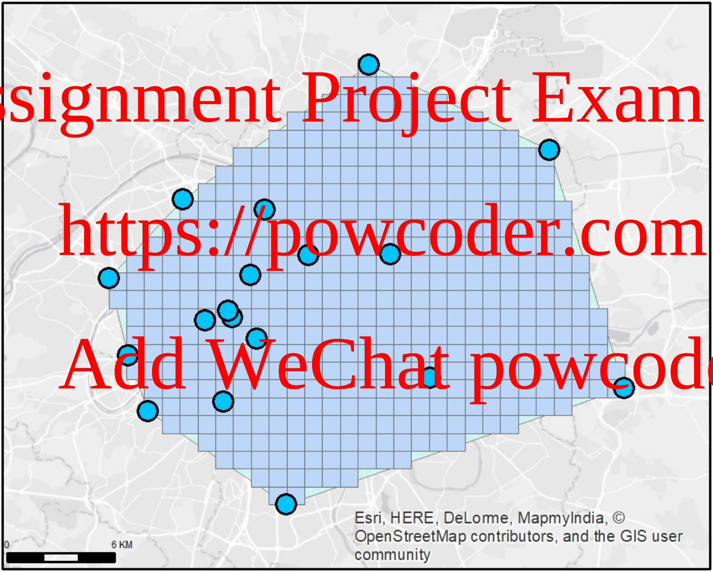
Paris - Prediction Points



Paris - Regular Grid of Points



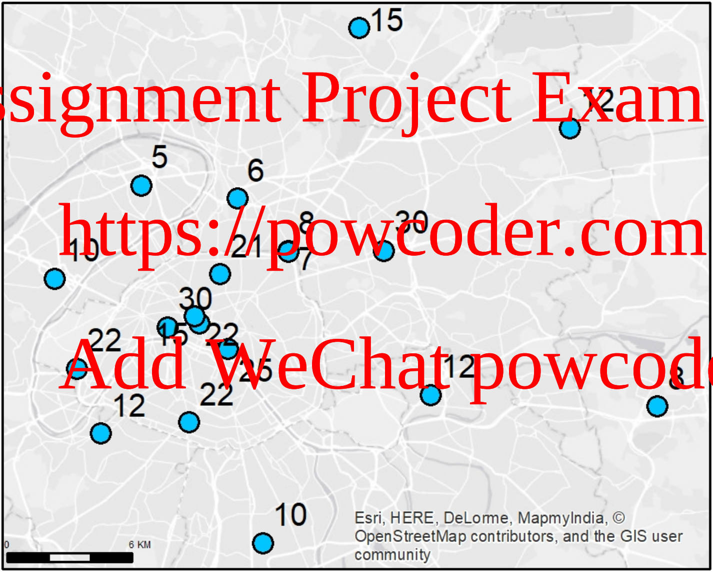
Paris - Regular Grid (Raster Output)



- Global vs local
  - Global interpolation takes into account all values
  - Local utilizes a moving window approach

- Exact vs approximate
  - Exact, value at a known location will remain
  - Approximate, values at a know location may change

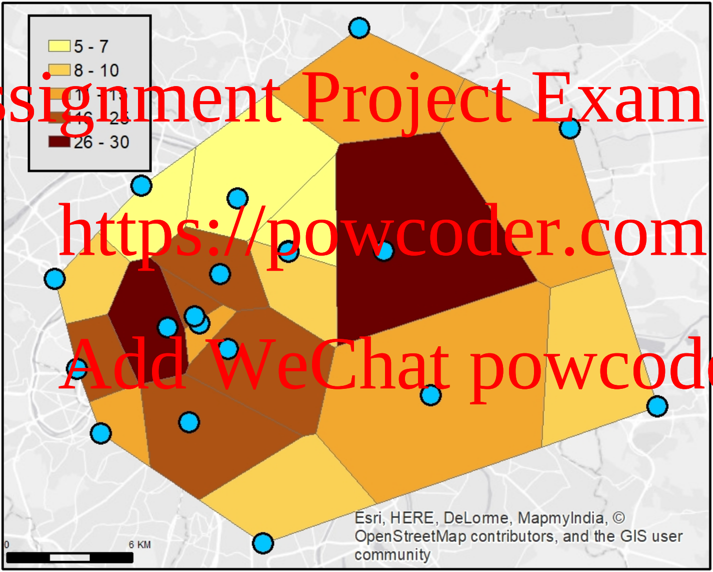
Paris - Pollution Values



Thiessen (Voronoi) polygons

- Individual areas of influence around each point of a set.
- Boundaries define the area that is nearest to each point relative to all other points.
- Mathematically defined by the perpendicular bisectors of the lines between all points.

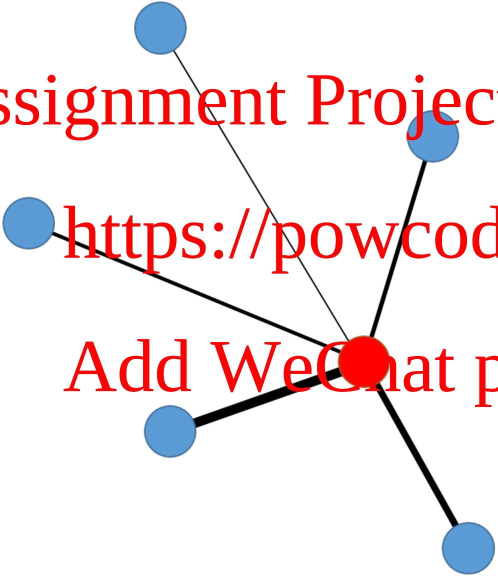
Paris Thiessen Polygons



- A very simplified interpolation technique
- Does not take into account:

  - surrounding values
  - more than one observation

Spatially Weighted Average



Inverse Distance Weighted Interpolation

- Exact Interpolator
  - Known locations retain their exact values

$$\hat{z}(x_0) = \frac{\sum_{i=1}^{n} z(x_i) \cdot d_{i0}^{-k}}{\sum_{i=1}^{n} d_{i0}^{-k}} \text{ if } d_{i0} \neq 0 \text{ for all } i$$

where, $d_{i0} \neq 0$ for some i, $\hat{z}(x_0) = x_i$; our prediction $\hat{z}(x_0)$ is based on a specified number of neighbours n, $i = 1, ..., n$

IDW Equation (Left of equals)

$$\hat{z}(x_0)$$

This is our predication for location x_0

IDW Equation (Bottom of Fraction)

$$\sum_{i=1}^{n} d_{i0}^{-k}$$

- n is the number of neighbours that will be considered
- d_{i0} the distance for each i to the prediction location
- $^{-k}$ is the exponent of distance
  - Often set to 2
  - Reduces the influence of points as they are further away

The sum of the distance measure (to the power of negative k) standardizes the function, as the total sum of distances will vary.

IDW Equation (Top of Fraction)

$$\sum_{i=1}^{n} z(x_i) \cdot d_{i0}^{-k}$$

Summing the product ($\cdot$) of the surrounding values and their spatial weight (d_{i0}^{-k})

IDW Equation (After Fraction)

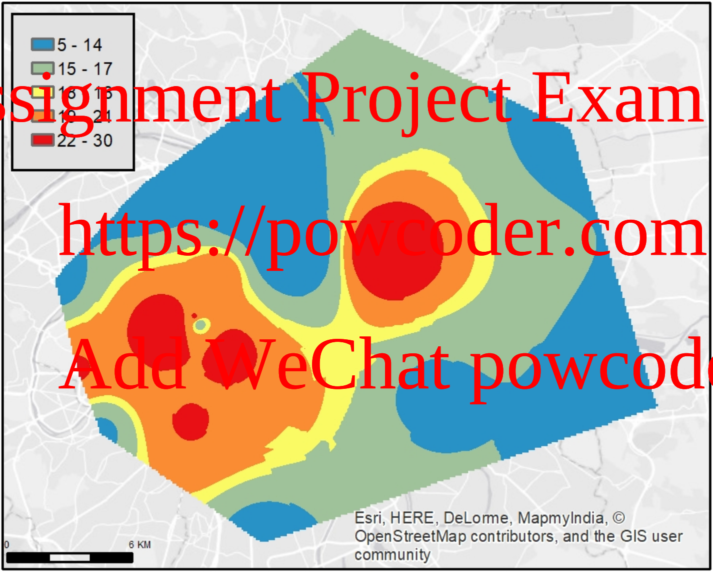
if $d_{i0} \neq 0$ for all i

- We will only conduct this calculation for prediction locations that do not match existing values
  - Exact interpolator

Selecting n and k

- n, number of neighbours that we will use in the calculation
  - Yes, we can use all neighbours
  - Dependent on number of cases
  - If you vary the n and the outcome is the same, it is less critical
- k, effects the weighting of distance
  - Smaller k, weights are higher for further points
  - Usually begin with $k = 2$
- Validate choices with cross-validation

Paris IDW Output



Cross-Validation

Types

- Exhaustive
  - Test all possible combinations to divide the original sample into a training and a validation set

- Non-exhaustive
  - Do not compute all ways of splitting the original sample.

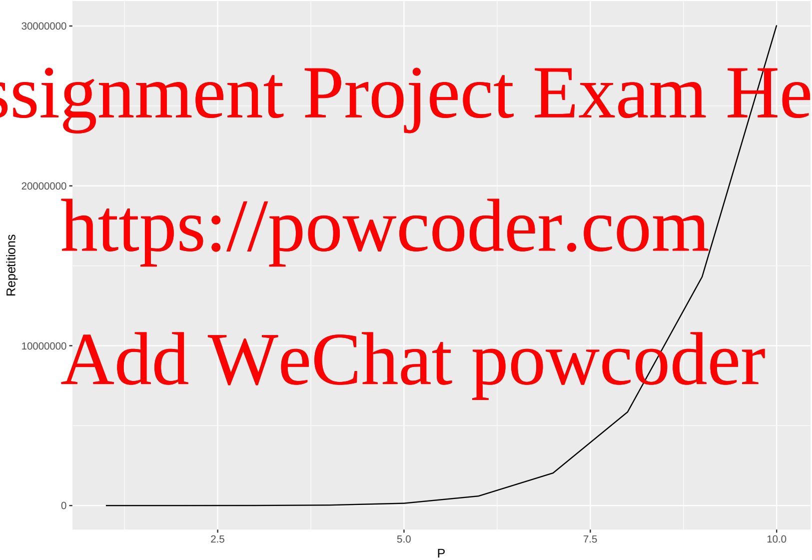
Exhaustive Cross-Validation

- Leave-one-out (LOOCV)
  - Remove a sample from the data
  - Fit the model
  - Compare estimated and actual value for the removed sample

- Leave-p-out (LpOCV)
  - Remove p samples
- LOOCV requires n repetitions
- LpOCV requires many more

$$\frac{n!}{p!(n-p)!}$$

LpOCV Repetitions (30 Samples)



Non-Exhaustive Cross-Validation

- Holdout Method
  - Split the data into a testing and a training set

- k-fold cross-validation
  - Original sample is randomly partitioned into k equal sized subsamples
  - A subsample is retained as the validation data
  - Remaining $k - 1$ subsamples are used as training data

- Repeated random sub-sampling validation (Monte Carlo)
  - Randomly split the dataset into training and testing data
  - Repeat n times
  - Repeated version of the holdout method

Measures of fit

Cross-validation requires a measure to assess the error.

- Mean squared error (MSE)
- Root mean squared error (RMSE)
  - Root mean square deviation (RMSD)
- Mean absolute deviation (MAD)

Cross Validation Output

Predicted Values

$$\hat{Y} = 2.3, 3.3, 2.5 \ldots n$$

Observed Values

$$Y = 2.2, 3.1, 1.7 \ldots n$$

Error

Error is observed minus predicted

$$Y_1 = 2.2 (Observed)$$
$$\hat{Y}_1 = 2.3 (Predicted)$$
$$\epsilon = -0.1$$

MSE

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Where,

$\hat{Y}$ is a vector of n predictions

Y is a vector of observed values

RMSE (RMSD)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}$$

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

Choosing a measure of fit

- MSE and RMSE are sensitive to outliers (extreme values)
  - Power of 2
- RMSE has the same units as the data being estimated.
- MAD is easily interpreted.
- RMSE or MSE is useful if large errors are more important than small errors