Spatio-temporal Approximation: Interpolation and Extrapolation Queries on Climate Data

Undergraduate Group 8
- Kayla Carter
- Nick Flowers
- Vanessa Sanders
- Jayden Stearns

CS-43016: Big Data Analytics

Introduction

- A major **global issue** is the rapid change and the degradation of Earth's atmosphere
- Many climatologists are limited to small-scale, controlled experiments in their particular region of the world
- **Estimation queries** on this big data could be used to supplement the amount of (or distance between) associated measurement devices

Spatio-temporal Climate Data

Potential Dimensions: Continent, Region, Latitude, Longitude, Day, Month, Year, Time (Hour, Minute, Second), etc.

Examples From Our Dataset:

Region	Country	State	City	DaysSinceC	AvgTemperature
Africa	Algeria		Algiers	0	64.2
Africa	Algeria		Algiers	1	49.4
Africa	Algeria		Algiers	2	48.8
Africa	Algeria		Algiers	3	46.4
Africa	Algeria		Algiers	4	47.9
Africa	Algeria		Algiers	5	48.7

Region	Country	State	City	Latitude	Longitude
North Ame	US	Iowa	Des Moi	41.5908	-93.6208
North Ame	US	Illinois	Peoria	40.7208	-89.6094
Africa	Algeria		Algiers	36.7717	3.0518
Africa	South Africa		Cape To	-33.9253	18.4239
North Ame	US	Alabama	Mobile	30.6889	-88.0448

Data Examples (continued)

Figure 1: CO_2 parts per million growth from previous year (e.g. In 2010, CO_2 increased by 2.3 PPM relative to 2009)

Year	CO2_PPM_Growth
1960	0.5
1970	1
1980	1.68
1990	1.2
2000	1.62
2010	2.3
2020	2.31

Figure 1

Figure 2: Earth's relative temperature by decade

Year	abTemp
1880	-0.16
1890	-0.35
1900	-0.09
1910	-0.43
1920	-0.27
1930	-0.16
1940	-0.13
1950	-0.17
1960	-0.03
1970	0.03
1980	0.26
1990	0.45
2000	0.39
2010	0.72
2020	1.02

Figure 2

Figure 3: Polar ice cap area (km^2) difference from Mean

Year	anomalyFromMean
1953	1.6
1958	1.2
1963	1.7
1968	2
1973	1.2
1978	0.9
1983	1
1988	0.9
1993	1
1998	0.4
2003	-4
2008	-2
2013	-2
2018	-2.3

Figure 3

Project Description

- Design an efficient querying algorithm to approximate values for unknown data points, using statistical estimation methods of Regression, Interpolation, and Extrapolation

- In our case, **Regression** involves creating a model for change in average temperature at a particular location over a specific time interval

- **Interpolation** and **Extrapolation**

Interpolation and Extrapolation

- **Interpolation** - estimating an unknown value within a sequence based upon other nearby surrounding values in the sequence
  - **Methods:** Function, Spline, Polynomial, Gaussian, Piecewise, Linear, etc.

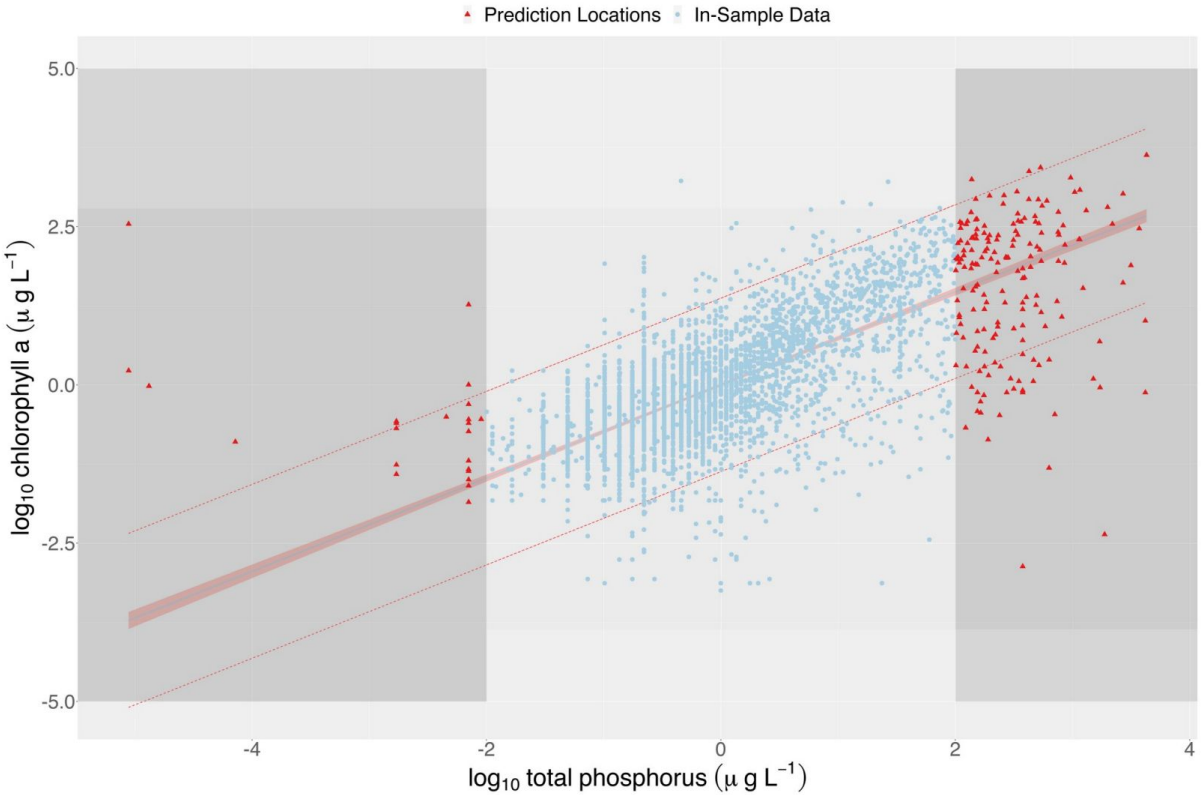
- **Extrapolation** - estimating an unknown value for a variable or function outside of the currently-recorded range
  - **Methods:** Function, French Curve, Conic, etc.

Similar Research

- **Extrapolation analysis** on lake nutrient and productivity variables across thousands of inland lakes in a geospatial and temporal database
  - Regression performed on **aggregate** of many datasets

Bartley, M. L., Hanks, E. M., Schliep, E. M., Soranno, P. A., & Wagner, T. (2019). Identifying and characterizing extrapolation in multivariate response data. PLoS One, 14(12). https://doi.org/10.1371/journal.pone.0225715

Similar Research (continued)



Problem Definition

Performs regression analysis across cities that account for **non-linear temperature gradients** across the landscape

We will perform a series of regression analyses. Each analysis will be for a city and the independent variable will be number of days since Jan 1st 1995. Our dependent variable will be Average Temperature.

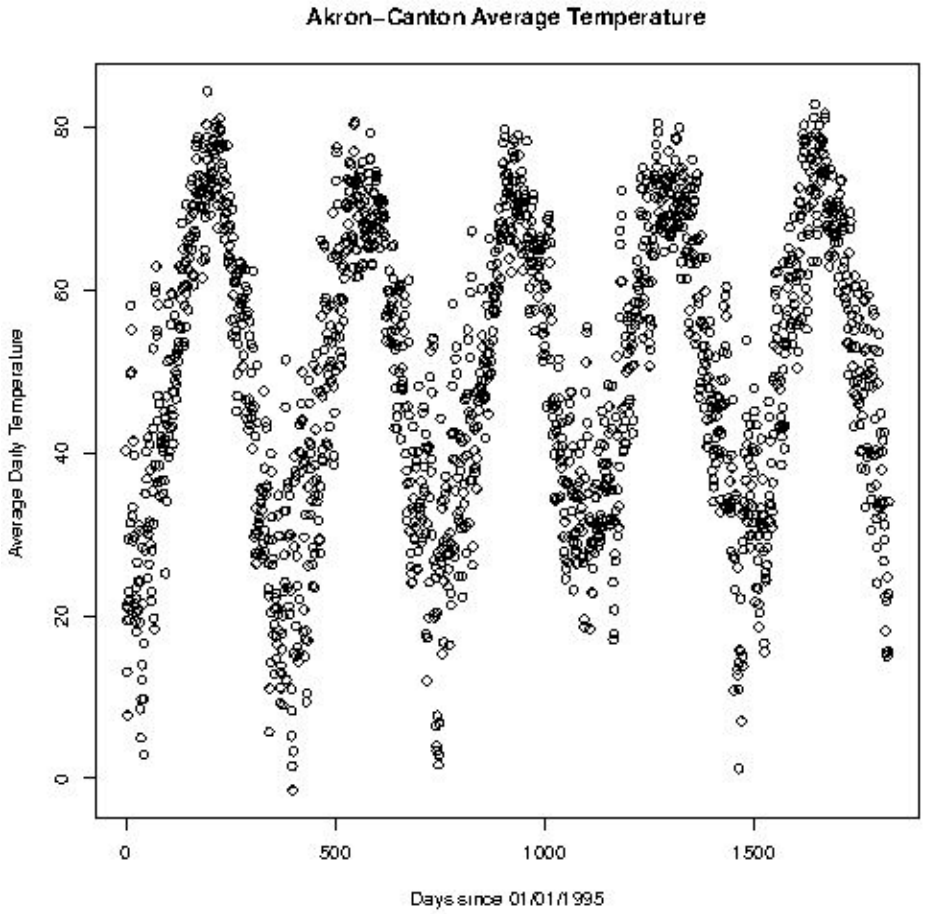
Solution Overview

- When an unknown data value is queried for, first determine the **locations which apply to this calculation:**

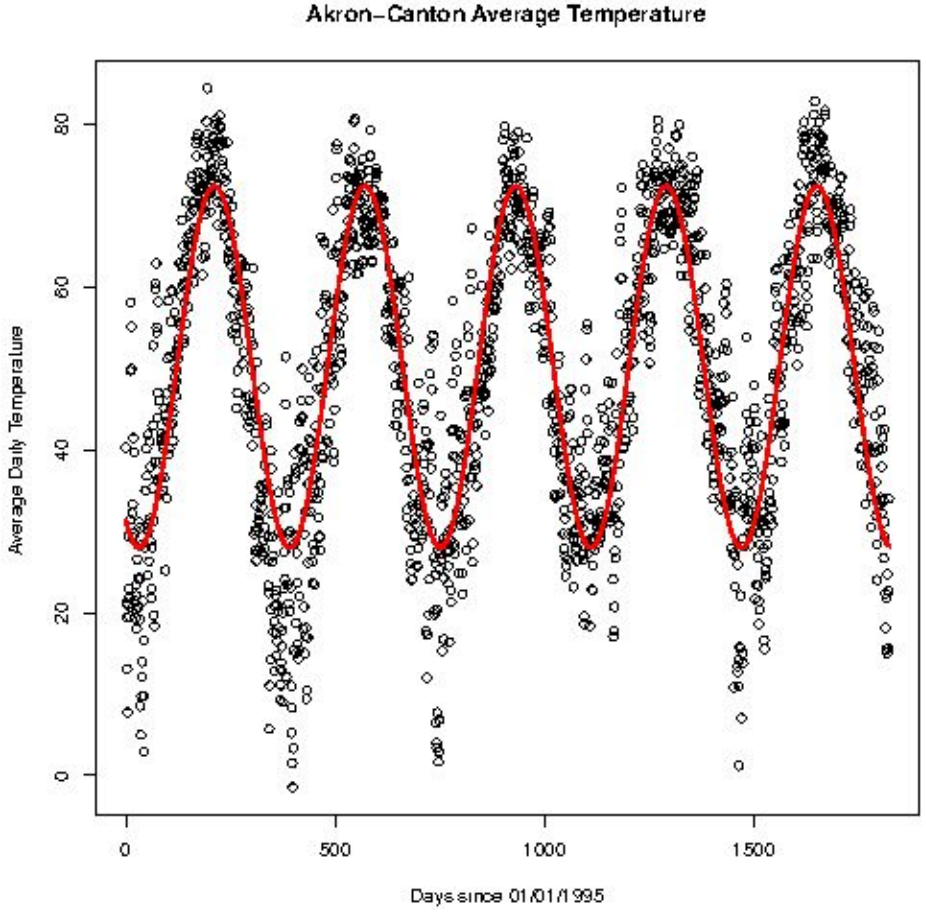
- If the data value is at a location for which **there are other known values** in time, calculate a regression model for the other known data and **approximate** the unknown queried value.

Solution Overview (continued)

- If the queried data value is **not** at a location that has other measurement values, perform a **nearest-neighbor query** on the desired location, retrieving the closest locations in order of proximity
  - Using these multiple regression models, a **weighted average** of their values by proximity to the desired location to produce an **approximation for the desired location and time**



Example: plot of one city's daily temperature over multiple years



Example: trigonometric regression fit to the dataset

Example Usage

- Approximate the daily temperature of Kent, Ohio on January 6, 1995. There is no measured data for Kent in the dataset, so other measured locations must be used
- The regression prediction for the only nearby cities (Akron-Canton and Youngstown) produce approximations of 27.85890 and 30.58912 degrees Fahrenheit, respectively
- Calculating the weighted average of these values by proximity to Kent, Ohio produces the value 28.76987 degrees Fahrenheit

Future Work

- Our work could be improved upon by studying different regression models that would provide even more accurate summaries and predictions of our data

- We could also expand this by performing multiple predictions to get closer nearest neighbors and utilizing that data to get more accurate predictions