John Bang
Stephanie Owyang

Github: https://github.com/gowrath/Project2_SOwyang_JBang

The primary dataset we intend to analyze is the Berkeley Earth dataset of global temperatures. We will analyze a subset of the data, focusing on temperatures in South American cities since 1900. We are choosing cities instead of countries because of the granularity of the data; in addition, cities may act as 'heat islands' where dense concentrations of pavement and buildings exacerbate heat waves. We are analyzing Lima, Santiago, and Sao Paulo.

Data structure

There are 1980 observations and 12 columns.

We will choose six variables:

- `year`
    - EX: 1900
    - This will help us determine the temperature reading of that year
- `month`
    - EX: 1 for January, 2 for February, etc
    - We can analyze the data for that particular month
- `monthly_anom`
    - EX: 1.344, 1.514, 1.636, -0.524
    - Using this variable, we can analyze the temperature increase or decrease for the measured month
- `monthly_unc`
    - EX: 0.826, 1.205, 0.208
    - "Uncertainties represent the 95% confidence interval for statistical noise and spatial undersampling effects. Such uncertainties are expected to account for the effects of random noise as well as random biases affecting station trends and random shifts in station baselines."
        - Shown as plus-minus on the monthly anomaly
- `oneyr_anom`
    - EX: 1.344, 1.514, 1.636, -0.524
    - Using this variable, we can analyze the temperature increase or decrease for the measured year
- `oneyr_anom`
    - "Uncertainties represent the 95% confidence interval for statistical noise and spatial undersampling effects. Such uncertainties are expected to account for the effects of random noise as well as random biases affecting station trends and random shifts in station baselines."
    - Shown as plus-minus on the yearly anomaly

The specific question at hand is what statistical properties can we glean from the data as we do our exploratory data analysis. The time series data set has columns for monthly, annual, five-year, to twenty-year anomalies (fluctuations) and uncertainty, which is a plus/minus range of the anomaly.

We expect to find insights on the temperature—for annual anomalies to increase over the last 100 years. We will also study the uncertainty values over time, to see if the fluctuations themselves change. We will compare the three cities using South American data as a baseline.

A supplemental database we will use is a population dataset of the cities. We may look at the economical development over time for the country as a whole over time (per capita income) and map that to temperature data. We may able to draw a correlation between human activity and climate change. (Such a correlation may be noisy and have several confounding variables, but we are in the exploratory phase for now).

Final Report

We want to explore a series of questions on climate data in South American cities:

1. Which cities have higher temperatures over time?
2. Do the temperature trends correlate to population trends?
3. Is there a particular season where the anomaly and uncertainty increased or decreased faster than others?
    a. To address seasonality, we should compare each season throughout each year, seeing if that follows the overall trend
    b. Are we seeing hotter summers and/or colder winters?
4. Can we discern large natural events or natural phenomena in the data, such as the pandemic?