It's Getting Hot in Here!

April 1, 2018

My first project[1] as part of my journey to "Digging Deeper" in the world of data to find insights hidden in plain sight. This project is part of the Data Analyst Nanodegree offered by Udacity and is the analysis of global and local temperature trends.

Table of Contents

[1] As this project was to explore basic data analysis process, Microsoft Excel was used. For future projects, I will be using Python or R due to accessibility to much better numerical and graphing libraries

Kings of Buzzwords

Due to the ubiquitous usage of words “Global Warming” and “Climate Change” in popular media over the past decade, my null hypothesis was that temperatures are increasing with time. A better goal was investigation of extent of increase in different regions over varying periods of time.

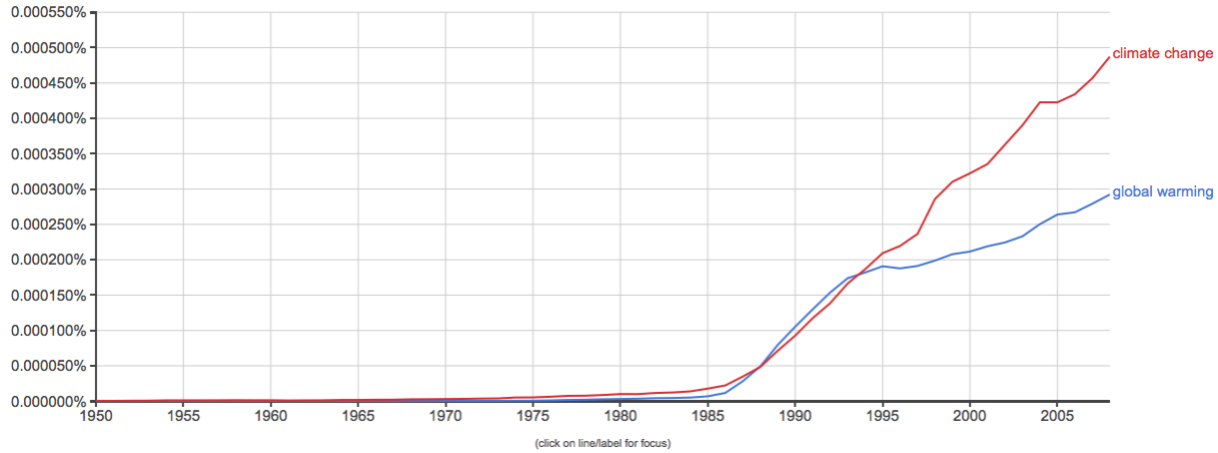


Figure 1 Popularity of "climate change", "global warming" in books according to Google Ngram Viewer

Data Extraction

Dataset 1

The first analysis was comparing Global and Toronto temperatures since it is my home city. Query to extract Toronto data:

```
SELECT *
FROM city_data
WHERE city='Toronto'
```

Query to extract global yearly temperatures:

```
SELECT *
FROM global_data
```

Dataset 2

I wanted to check whether the rate of temperature change differed depending on the climate of a city. My hypothesis was that the rate of increase is higher in colder climates due to feedback loops where melting ice further increases rate of melting.

Note: All temperature values in report are in Celsius.

- Lowest average temperature city

```
SELECT city, avg(avg_temp) AS combined_Avg_Temp
FROM city_data
GROUP BY city
ORDER BY combined_Avg_Temp
```

- Highest average temperature city

```
SELECT city, avg(avg_temp) AS combined_Avg_Temp
FROM city_data
GROUP BY city
ORDER BY combined_Avg_Temp DESC
```

- Median average temperature city (LIMIT 173 as 345 total city rows)

```
SELECT city, country, avg(avg_temp) AS combined_Avg_Temp
FROM city_data
GROUP BY country, city
ORDER BY combined_Avg_Temp
LIMIT 173
```

Chosen Cities:

- Cold: Ulaanbaatur, Mongolia (Average Temp: -3.67)
- Hot: Bangkok, Thailand despite Khartoum, Sudan having the highest average temperature due to better data quality.
    - Bangkok, Thailand: Data available since 1816 (Average temp: 27.15)
    - Khartoum, Sudan: Data available since 1859 with missing data
- Median: Athens (Average Temp: 17.42)
    - Not median city but better data quality

Data Cleanup

Missing Values

- Bangkok: Estimated missing values for years 1824 – 1832 using Simple Moving Averages with lag of 8 years.[2]
- Missing data for 2014-15 for all cities *except* global data. Estimated using moving average with shorter lag of 5 years due to increasing frequency of higher temperatures in recent years (details later).

Data Smoothing

Issues with raw data:

- Measurement Error: Equipment used for variate (temperature) measurement was inaccurate in 19th century
- High variability in year-to-year temperatures which is undesirable as I'm interested in long term trends

Solution: Used simple moving average with variable lag factors to smooth data for plotting

Moving Averages: Choosing a Lag Factor

The lag factor chosen has a significant impact on trends observed: a high lag causes unusually smooth graphs diminishing any trends and a low lag causes highly variable graphs exaggerating year-to-year temperature jumps.

To decide on a lag factor for different time periods, I researched CO_2 levels over the past three centuries. CO_2 is a greenhouse gas which prevents the escape of heat from the earth's environment and has drastically increased in concentration due to the increased burning of fossil fuels. Thus, it is a good indicator of increasing overall temperatures.

The measurement of yearly CO_2 levels began in 1959 at Mauna Lao, Hawaii which shows increasing levels every year since 1959.[3] Thus, to capture the increasing effect of CO_2 in the atmosphere, I used a lower lag factor of 7 beginning 1950. Furthermore, I analyzed the annual rate of growth of CO_2 at Mauna Loa which is shown in Figure 2.

---

[2] Simple Moving Average is one of the simplest forecasting methods where the next value is the mean of previous n periods where n is lag factor.

Example: Temperature in year 11 with lag factor of 5 is $\sum_{i=6}^{10} t_i / 5$

[3] Increasing yearly CO_2 concentration: https://www.esrl.noaa.gov/gmd/ccgg/trends/full.html

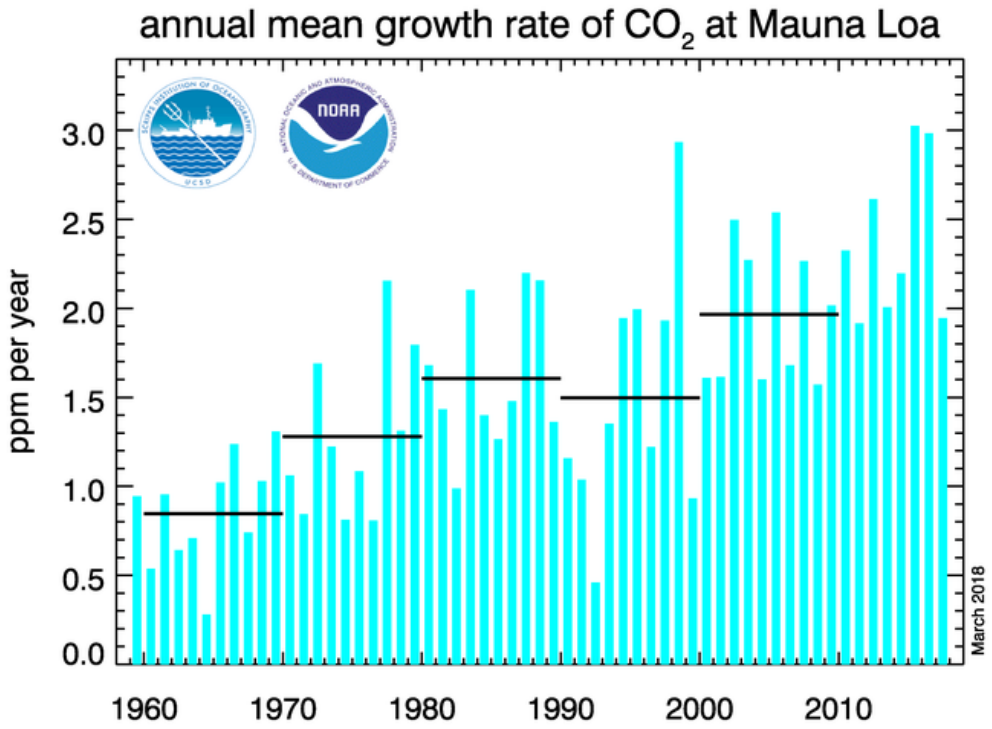


Figure 2: Annual rate of growth of Carbon Dioxide in the atmosphere

Based on these factors, I chose the following lag factors

TIME PERIOD	LAG	REASON
PRE-1950	10 years	Eliminated year-to-year variability without losing overall trend
1950 – 1995	7 years	Increased levels of CO_2 levels due to industrialization
1995 – 2005	5 years	Even higher annual growth of CO_2 levels in late 20th century
2005 – 2015	3 years	Extremely high increase in CO_2 levels growth

Lag Factor Exceptions

The following table shows basic descriptive statistics of different climate regions.

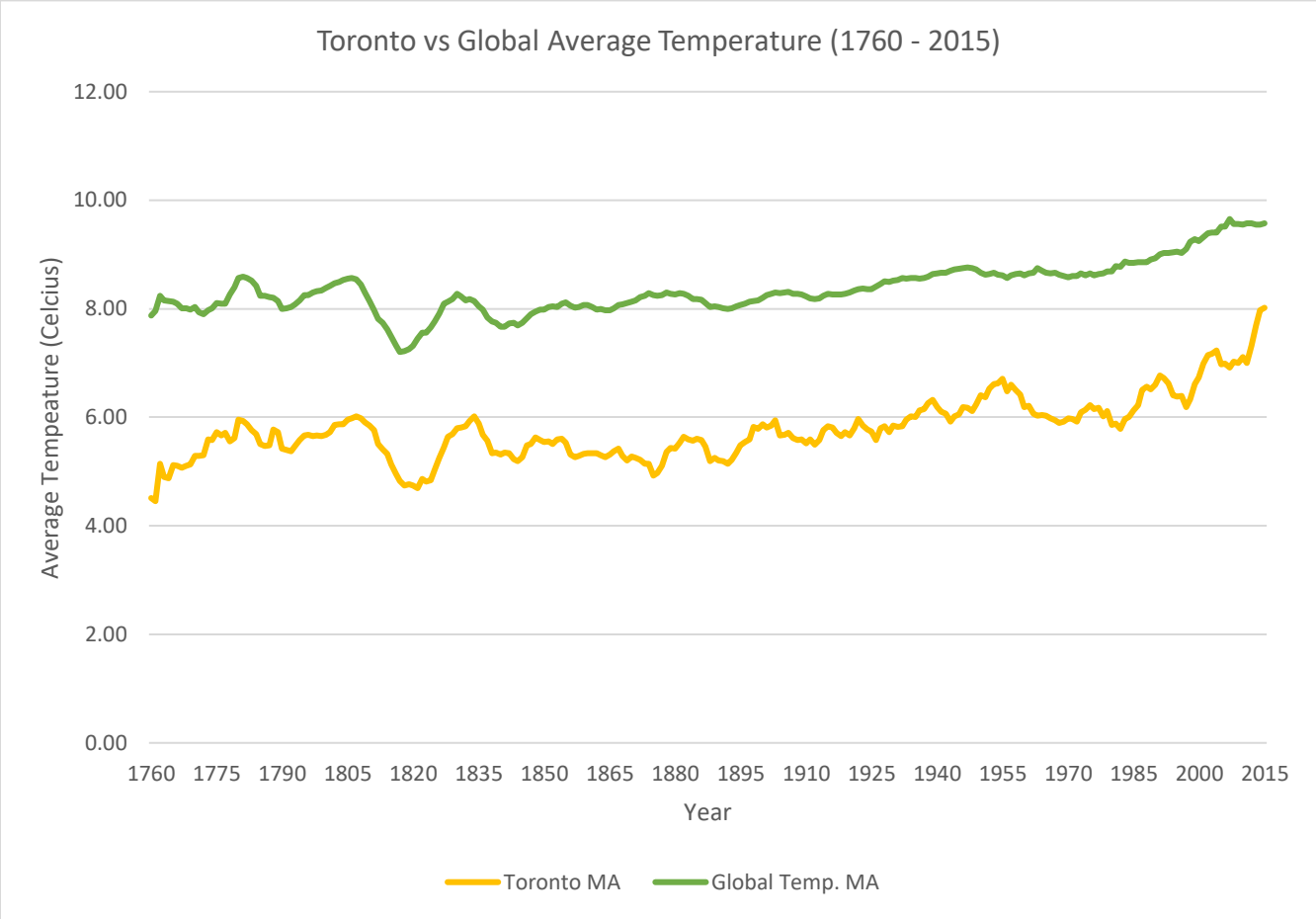
	GLOBAL	HOT	AVERAGE	COLD	TORONTO
STD. DEVIATION	0.58	0.55	0.57	1.02	1.01
MEAN TEMP. (CELSIUS)	8.37	27.15	17.43	1.06	5.78

The standard deviation of temperatures in colder regions is higher than other climates. To account for the higher variability, the lags chosen were higher:

- 1950 – 2005: Lag of 7 years
- 2005 – 2015: Lag of 5 years

Exploratory Analysis: There was a Little Ice Age!

After smoothing the data with appropriate lag factors, I plotted an exploratory chart of Toronto vs Global temperature to get an initial sense of data.

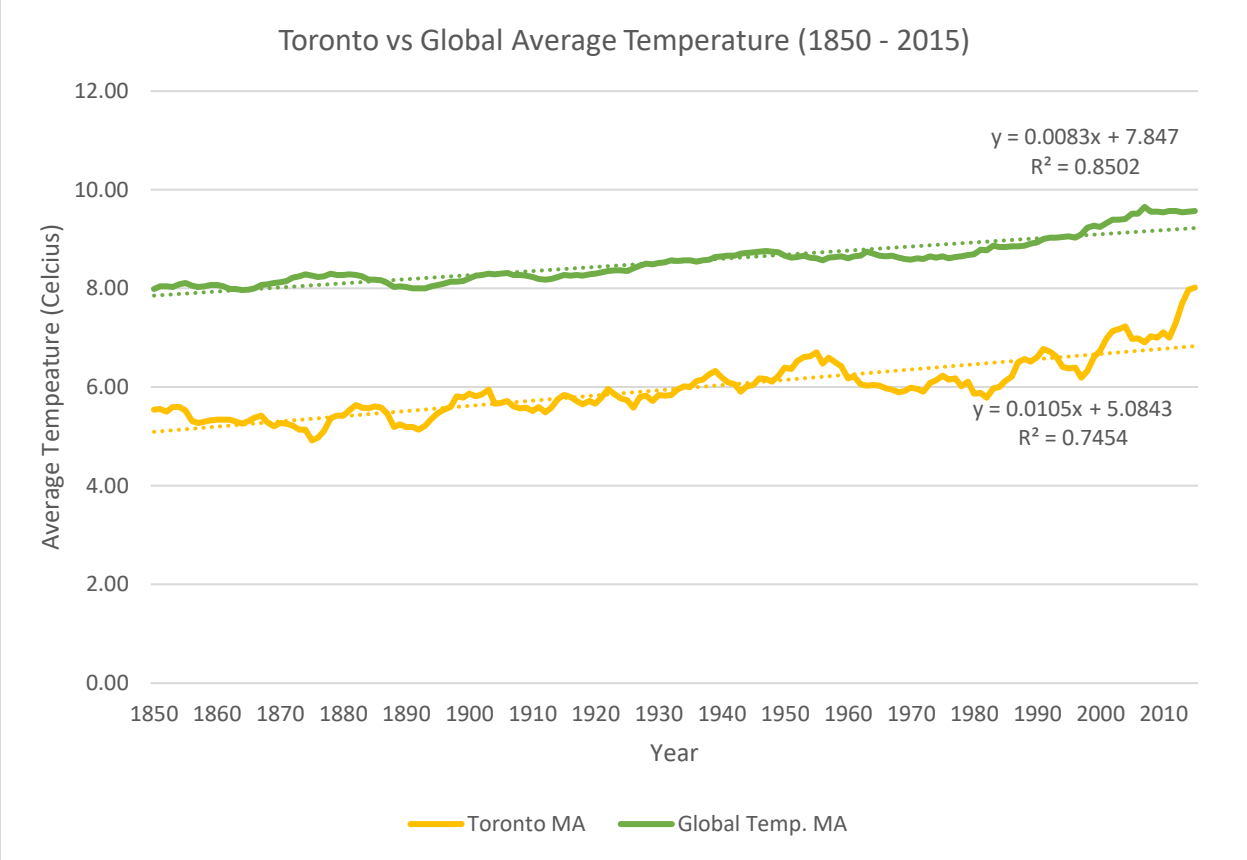


I noticed a huge drop in temperatures at the start of 19th century up until 1850. Upon investigation I found that the period from 1400 – 1850 had a Little Ice Age (LIA)[4] during which climate around the world dropped. Furthermore, I was interested in analyzing the impact of human industrialization which exponentially increased after 1870 due to the Second Industrial Revolution.[5] Due to Little Ice Age and minimal human advancement pre-1850, I decided to exclude all temperature data before 1850.

[4] Little Ice Age (1400-1850): https://www.britannica.com/science/Little-Ice-Age
[5] Industrial Revolution Timeline: https://www.ducksters.com/history/us_1800s/timeline_industrial_revolution.php

Time Period Chart Analysis

Period 1: 1850 – 2015



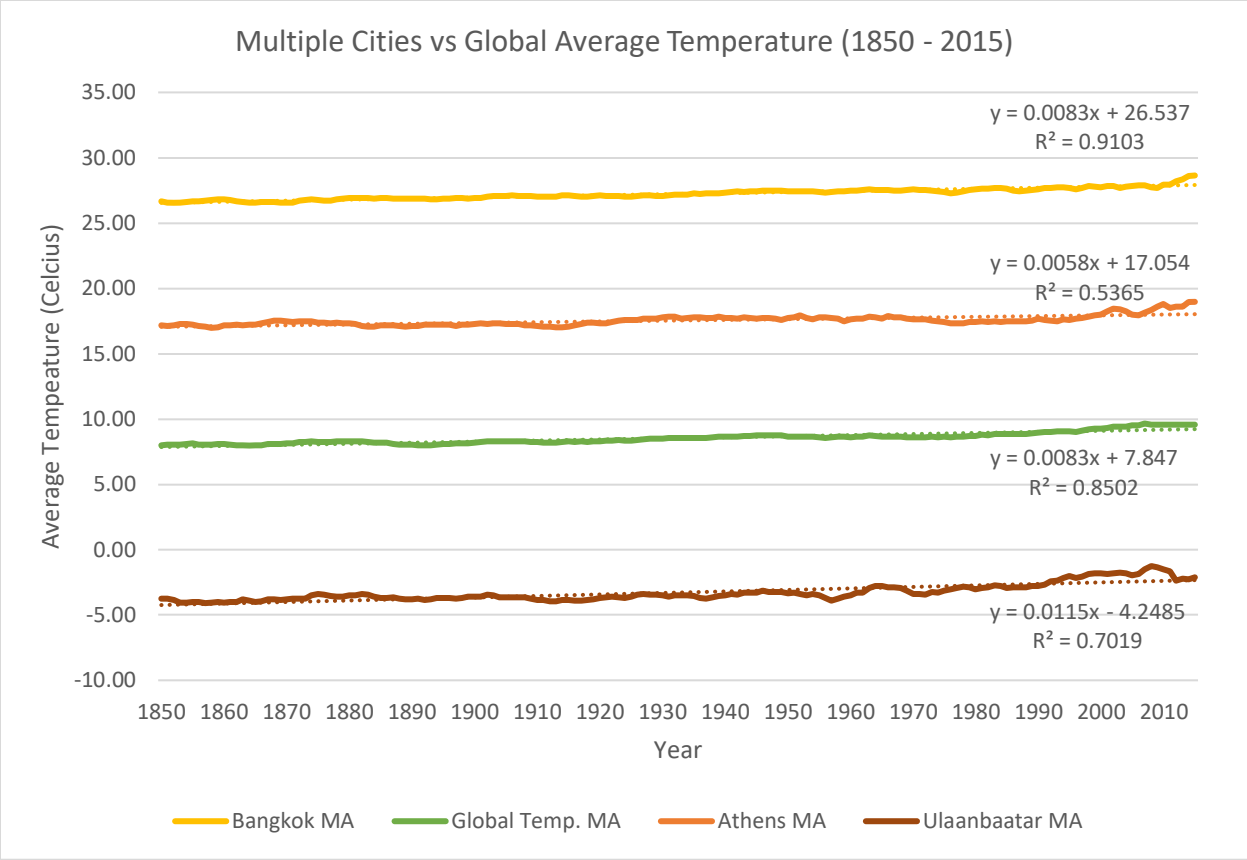
Observations: The increase in Toronto was greater than global (coefficients are 0.0105 vs 0.0083 respectively). Since Excel does not automatically calculate other linear regression statistics, I manually calculated parameters which are more reliable than R-squared value. Refer to footnotes for non-statistics definition of the terms.

	R-squared[6]	*Adj. R-squared*[7]	*Std. Error of Regression*[8]
Global	0.85	0.85	39%
Toronto	0.75	0.74	50%

Even though the R-squared value is relatively high for both models, the standard error of regression is relatively high for Toronto model.
The regression line was significantly below data points in the 21st century suggesting further investigation using smaller time periods was necessary.

---

[6] R-squared: Measure of how close the data are to the linear regression line (Higher is *usually* better)
[7] Adj. R-squared: Unbiased version of R-squared which accounts for number of samples (years) and number of variables predicted (only 1: temperature)
[8] Std. Error of Regression: How wrong the regression model is on average in units of temperature (Lower is better)



Observations: Due to the large range of years, the linear models again had high standard errors of regression. The coefficients of the models again showed greater increases in temperature for colder climates. However, the data had to be analyzed in smaller time period buckets to better examine recent trends.

Period 1: Residual Plot Analysis

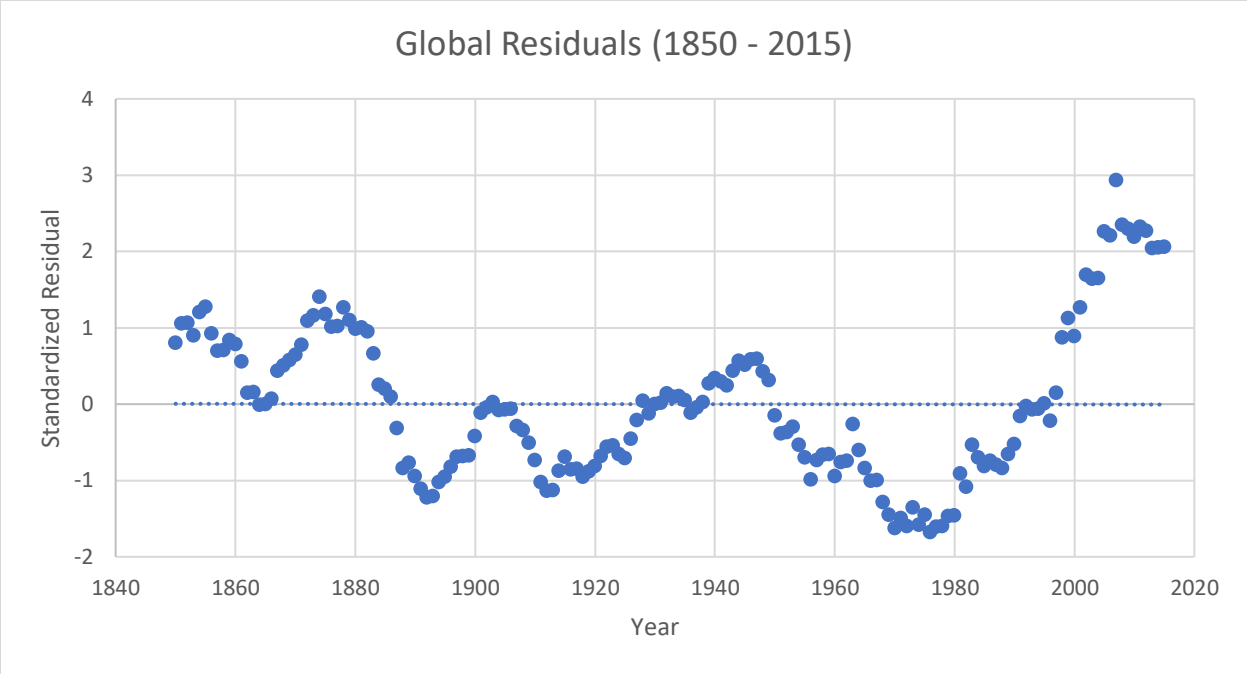
For those unfamiliar with residuals: A residual is the difference between an actual data point (eg: temperature measured in 1880) and the temperature on the regression line for the same year. Hence, a residual plot contains residuals for all years plotted against the independent variable (Year). Here are the basic components of a valid regression model:

Response (Temperature) = [Constant + Predictor(Year)] + Error

Response = Deterministic + Error

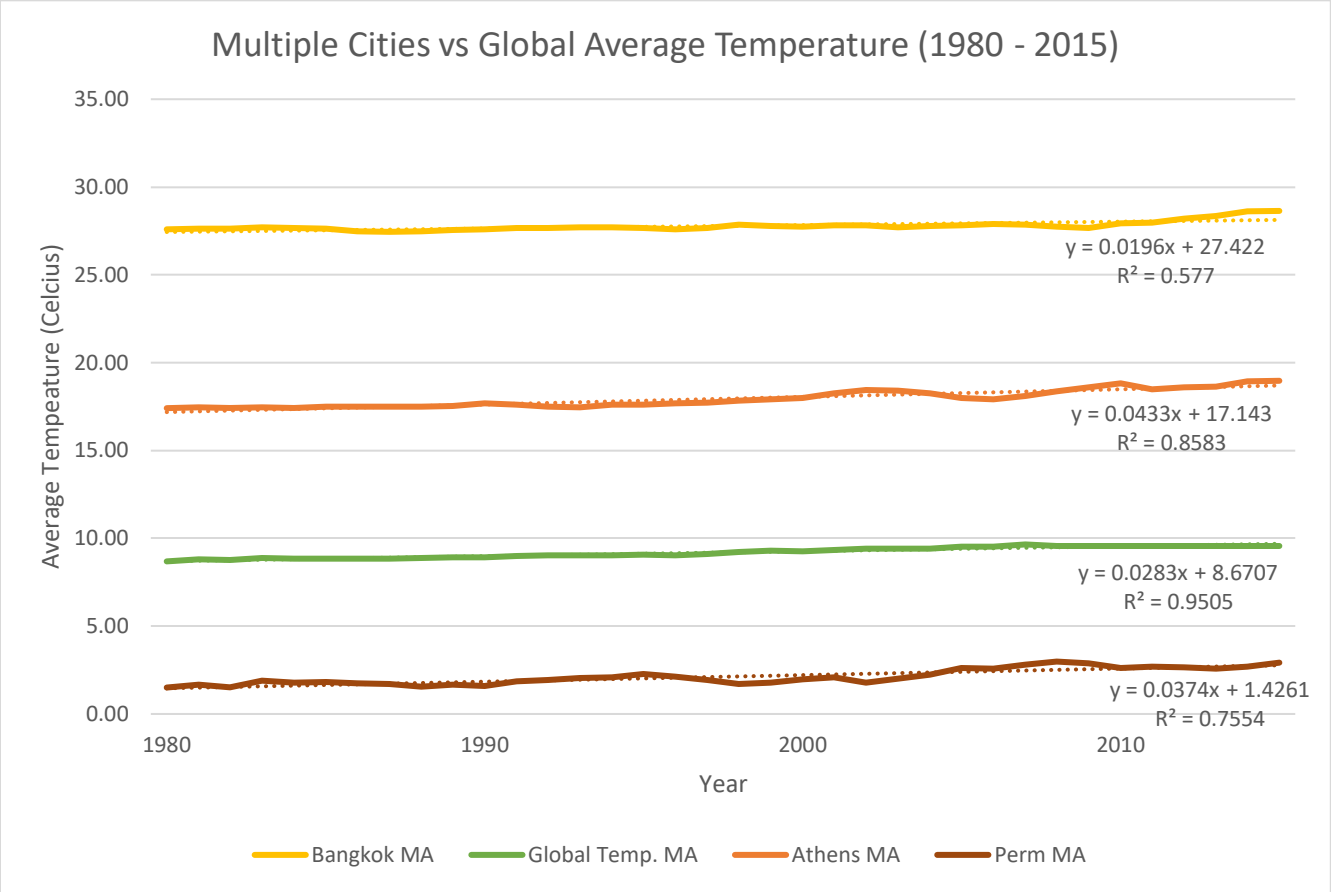
Using residual plots, you can assess whether the observed residuals (errors) are randomly distributed. The errors obtained must not have any visible trends and should be evenly distributed around the x-axis. The presence of visible trends implies that the regression model is not capturing all deterministic information which is 'leaking' into the error.[9]

[9] Why residual plot analysis is necessary? : http://blog.minitab.com/blog/adventures-in-statistics-2/why-you-need-to-check-your-residual-plots-for-regression-analysis



This residual plot is showing crazy trends and is definitely not randomly distributed. This might be due to the large time period being captured by the model. Next, I looked at smaller time periods to obtain a better model, specifically looking at periods when the carbon dioxide levels started increasing at a higher rate.

Period 2: 1980 – 2015

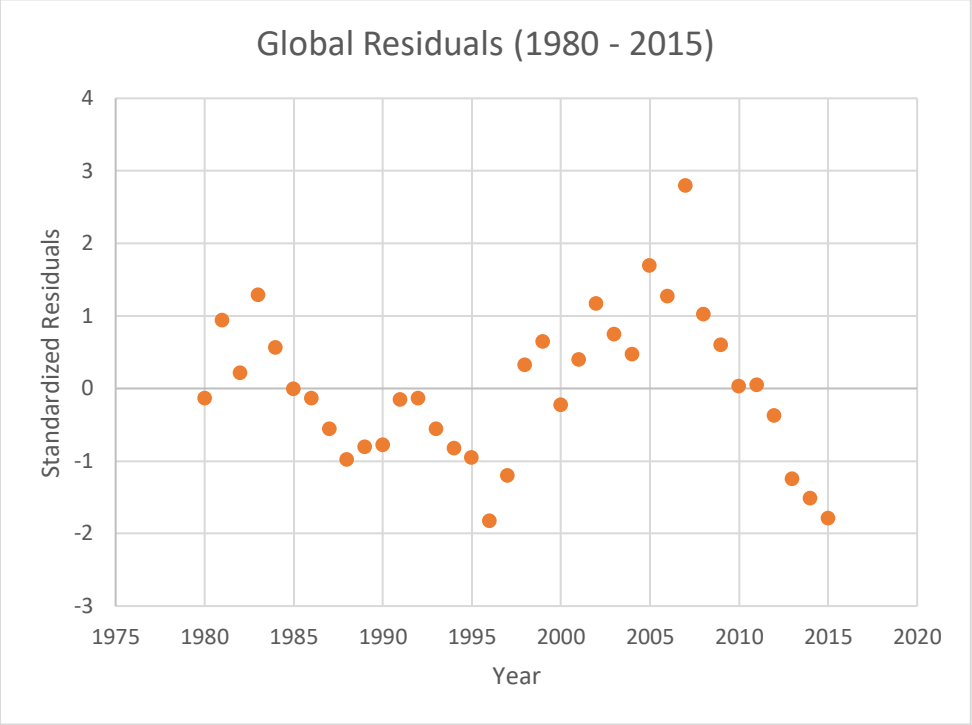


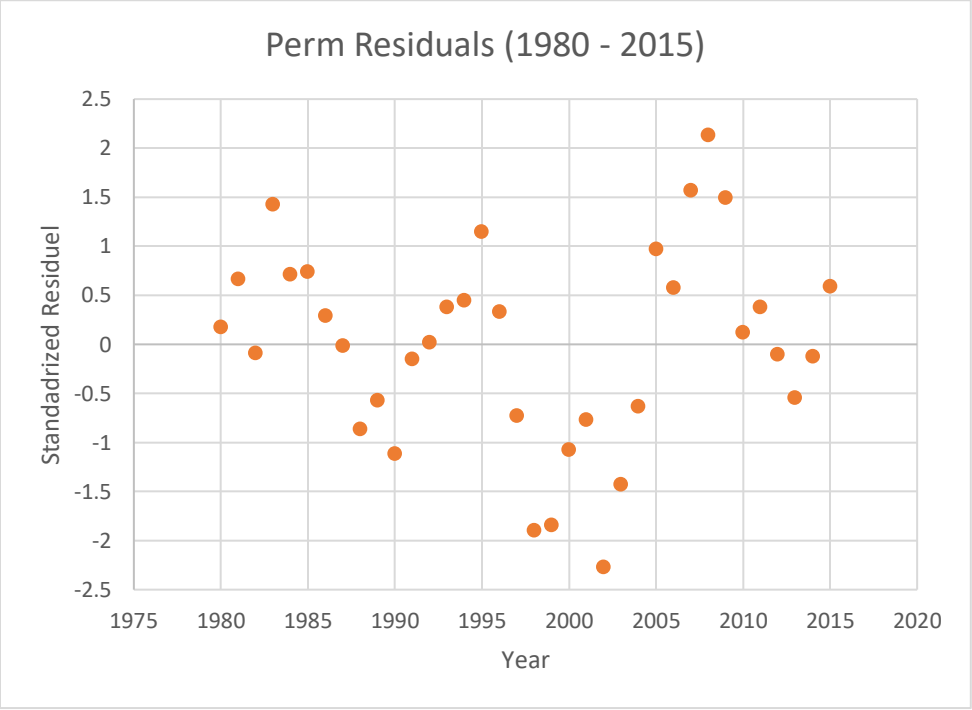
The coefficient values are significantly higher for this time period compared to larger period as highlighted below.

	1850-2015	*1980-2015*	*Coefficient Increase (%)*
Bangkok	0.0083	0.0196	236%
Athens	0.0058	0.0433	747%
Perm	0.0112	0.0374	334%
Global	0.0083	0.0283	341%

Another interesting observation is that the rate of increase for milder climate (Athens) which was below average in pervious graph is comparable to increases in colder climate (Perm). The increase is slower in hot climates (Bangkok) which is consistent with the previous observations. Overall, the rate of increase of temperature has increased considerably in past few decades.

Period 2: Residual Plot Analysis

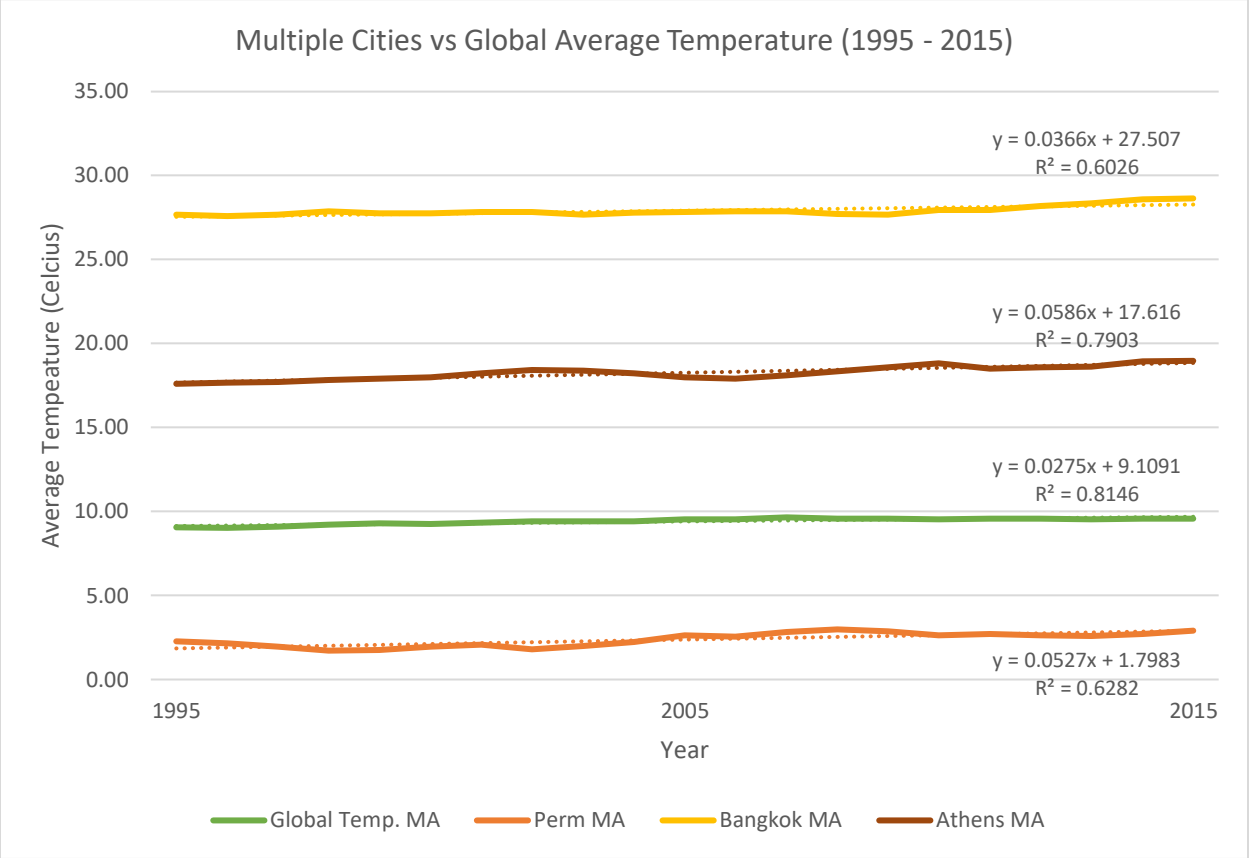




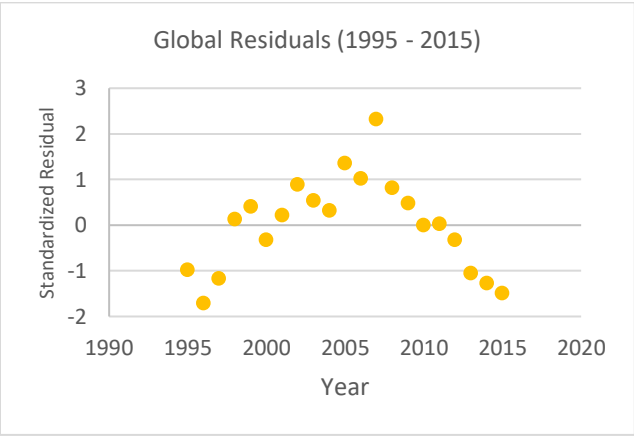
The residual plots for this time period are still not perfectly randomly distributed along the x-axis but the trends present are significantly minor compared to residual plots for larger period. It shows that the regression model has less deterministic information “leakage”.

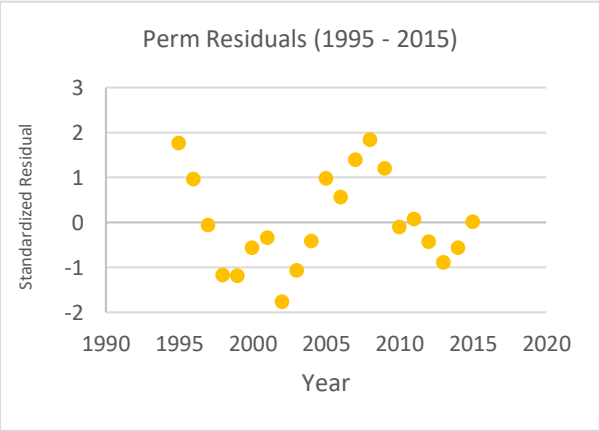
Period 3: 1995 – 2015

I wanted to analyze an even smaller time period which had the highest growth rates of carbon dioxide in the atmosphere.



Unsurprisingly, the rate of increase corresponded with the higher increases in growth rates of carbon dioxide as the coefficients for the cities further increased. However, the global coefficient had minor fluctuations.





The residual plots had stronger visible trends for this period implying the leak of more predictor information.

Concluding Thoughts

The trends in the global temperature and cities is clear with evidence of global warming in all climates. The rate of increase is higher in mild to colder climates with the overall rate of warming drastically increasing over the past few decades.

Limitations

The exact rates of temperature increase must not be used in other models for any purpose due to the missing deterministic information in all regression models. This was highlighted by the residual plots which had visible trends proving leakage of predictor information.

This project also shows the shortcomings of Microsoft Excel's linear trend function. It is useful to obtain the general trends of a phenomenon but the obtained models must be thoroughly validated before being applied at other places.