Data Story

Robert Munch

November 5, 2018

# Introduction

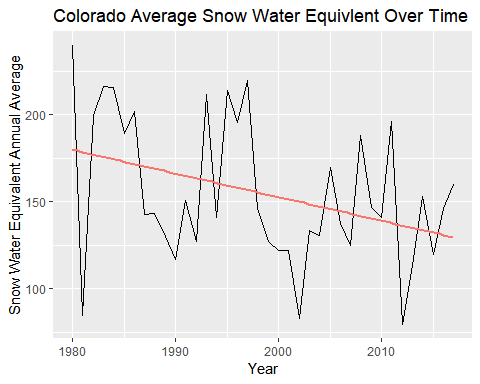
Colorado, in the mountain west region of the United States has a section of the central rocky mountains running through it north to south splitting the state in half, and giving it the climate and availability of some of the best known ski areas in the world. The question I wanted to answer is whether the state of Colorado will be feeling the affects of a rising CO2 concentration in the atmosphere in the form of lower snowfall, a poor outcome if your reputation relies on freezing temperatures and moisture in the atmosphere. Also, because the state generates revenue from the ski tourism industry, I wanted to know what affect economic impact lower seasonal snowfall has on the state.

# Snow Water Equivalency Data

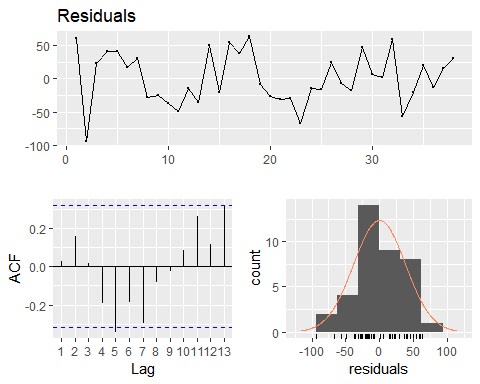
Using data gathered from the National Resource Conservation Service (NRCS) Snow Telemetry (SNOTEL) stations, the average snow water equivalent (SWE) in millimeters for the year was gathered. Data was gathered from every available station, a total of 113 stations in total, from around the state. The data was recorded giving a sum of the snowfall at the eand of each month, and did not include the amount of SWE that was already on the ground.

Becasue of the seasonality of snowfall, summer months almost always gave zero SWE values, making the data incoherant, therefore annual average of SWE was used because it gave cleaner data points, and still was able to show enough differentiation over time to show strong trends to the data.

Also, becasue of a natural phenomenon, there is an annual cycle known as El Nino and La Nina years which show a distinctive freequency in the data.

 Figure 1 This plot shows the data points and trend for annual mean SWE data as well as a linear fit.

##   
## Call:  
## lm(formula = SWE.mm.Avg ~ Year, data = SWE.year.agg)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -93.810 -26.661 -8.144 30.094 62.831   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2867.2886 1152.7831 2.487 0.0176 \*  
## Year -1.3573 0.5768 -2.353 0.0242 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 38.99 on 36 degrees of freedom  
## Multiple R-squared: 0.1333, Adjusted R-squared: 0.1092   
## F-statistic: 5.537 on 1 and 36 DF, p-value: 0.0242

Figure 2 This plot shows the data points and trend for annual mean SWE data as well as a linear fit. 

##   
## Breusch-Godfrey test for serial correlation of order up to 7  
##   
## data: Residuals  
## LM test = 10.276, df = 7, p-value = 0.1734

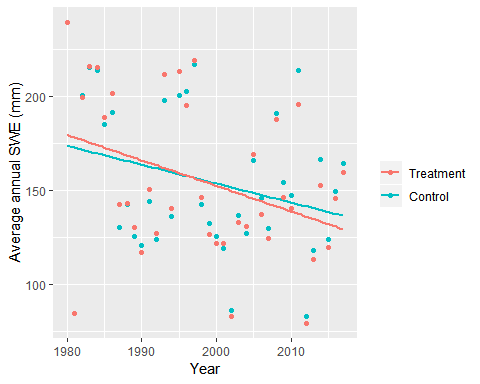
Figure 3 This shows residual data for the linear model of SWE annual mean over time

As seen in Figure 1, the El Nino and La Nina year cycles are significant, giving a farly consistant change in snowfall over the course of the 27 years, but even with these effects, there is still a downward trend to the annual average data, showing that even with the cyclical nature of this data, annual snowfall is progressivly getting lower over time.

As seen in the analysis of the frequency data generated from the linear model, the cyclical nature shows up most strongly on the ACF plot, and the histogram is fairly normally distributed.

Because the year 1981 seems to be especially low, and the available data is less complete earlier in the dataset, I wanted to make sure the variation was not due to certain monitoring stations only used in earlier data collection that had a higher chance of collecting low data.

Using only sites available in 1981, I averaged annual SWE data and compared it to the original values. The data is below with the annual average SWE data for all stations displayed in blue and the data for only the stations available in 1981 displayed in red.

 Figure 3

##   
## Pearson's Chi-squared test  
##   
## data: PrecipData.outlier$SWE.mm.Avg.outlier and PrecipData.outlier$SWE.mm.Avg  
## X-squared = 1406, df = 1369, p-value = 0.2378

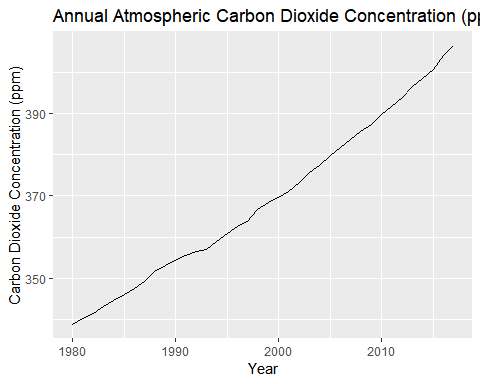
Based on the chart, it is obvious the plot with only sites available in 1981 is very similar to data using all available sites. Also, based on the p value from the chi-squared test, using standard alpha value of 0.05, we cannot confirm the null hypothesis. Therefore the data do show that the low SWE value in 1981 is highly unlikely to be due to selecting stations that have a lower average SWE.

# Atmospheric Carbon Dioxide concentration.

Using Data gathered from NOAA Earth System Research Laboratory, average monthly concentration in ppm is displayed over time.

Data was used from the Interplated column, as this column replaces missing values with an expected value.There was one missing value point in MArch of 1984, using the interpolated column gave me the ability to have a more complete data set and to be able to do less clean up of the data set to have it meet the other sets.

Becasue the line is very linear, no linear model was generated.

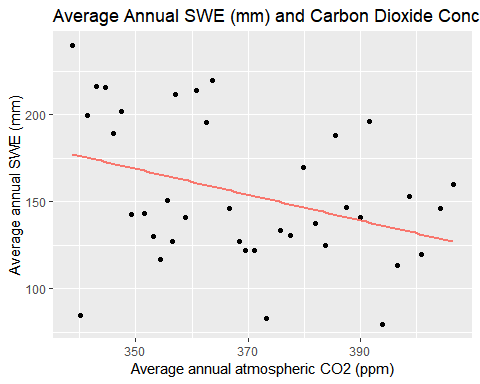


##   
## Call:  
## lm(formula = co2.ppm.Avg ~ Year, data = co2.year.agg)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.7439 -1.5450 -0.0066 1.3493 4.2646   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.210e+03 5.462e+01 -58.76 <2e-16 \*\*\*  
## Year 1.791e+00 2.733e-02 65.52 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.847 on 36 degrees of freedom  
## Multiple R-squared: 0.9917, Adjusted R-squared: 0.9915   
## F-statistic: 4293 on 1 and 36 DF, p-value: < 2.2e-16

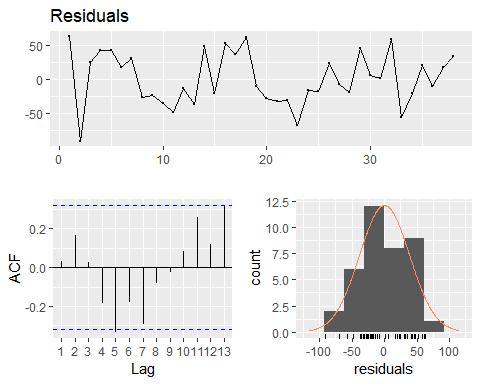
Figure 4

Based on these linear models, I expect that concentration to follow a linear equation of meaning that if this trend were to continue, in another 20 years with the current level of accumulation, the ppm of in the atmosphere will reach 440 ppm.

Next we will look at what this means in context of snowfall.



##   
## Call:  
## lm(formula = SWE.mm.Avg ~ co2.ppm.Avg, data = SWE.co2.year.agg)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -91.591 -26.043 -9.087 32.261 62.468   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 428.2543 118.9057 3.602 0.000947 \*\*\*  
## co2.ppm.Avg -0.7412 0.3216 -2.304 0.027077 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 39.1 on 36 degrees of freedom  
## Multiple R-squared: 0.1285, Adjusted R-squared: 0.1043   
## F-statistic: 5.31 on 1 and 36 DF, p-value: 0.02708



##   
## Breusch-Godfrey test for serial correlation of order up to 7  
##   
## data: Residuals  
## LM test = 9.7898, df = 7, p-value = 0.2008

Figure 5

This figure compares the average annual SWE data for all sites across CO, and the average annual concntration in parts per million. This has been fitted with a linear line of best fit and data showing residual data. Based on this plot, it shows there is a negative linear coorelation between concentration and average annual SWE in Colorado. This means that as concentration increases, SWE decreases.

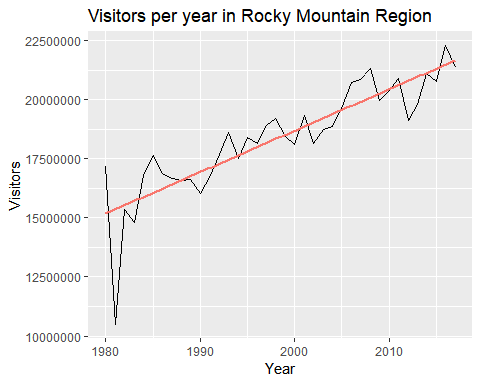
Similarly to the previous plot showing only SWE data over time, the histogram generated from the residuals is fairly normally distributed.

Based on the linear model of past average annual snowfall and concentration, I expect the SWE to follow a linear model of meaning that for every ppm increase of in the atmosphere, there will be a reduction of 0.7412 mm of SWE in Colorado.

Lets apply this to the 20 year increase we expect in ppm. With a level of 440 ppm of in the atmosphere, I predict an average snowfall of 102 mm. That is a number only currently achieved during the worst of the drops curently in the data.

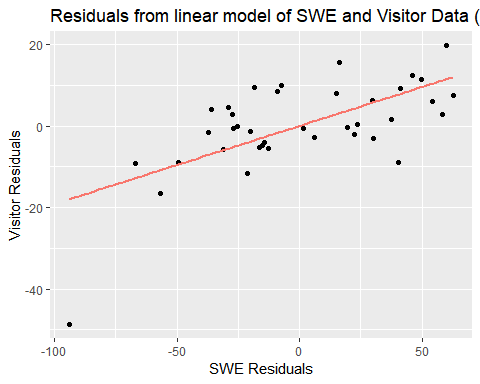
# Visitor Data

Because there was no publicly available data available for only the state of Colorado or to specific ski resorts, data made available by the National Ski Area Association which gives visitor numbers by region had to be used. Because Colorado is in the Rocky Mountain region and visitor data averall should coorelate to Colorado quite fairly, data was manually scraped from a reprt released by the National Ski Area Association. Visitor data is used to see if there is a correlation with visitors to SWE data. Thus showing if lower snowfall causes a decrease in visitors to ski areas.

 Figure 6

This plot showing the number of visitors each year in the Rocky Mountain region, shows that there is an overall increase in the number of visitors to ski areas, however there is shifts in the residual amounts for each year.

Because the visitors will continulously be going up, and there are many other factors that influence why more people seem to be skiing each year, I plotted the residuals from the linear model and compared those to the residuals from the SWE model to compare if lower than average snowfall results in lower that average visitors. This should give a good indication whether lower than average snowfall also generates lower than average visitors to an area.

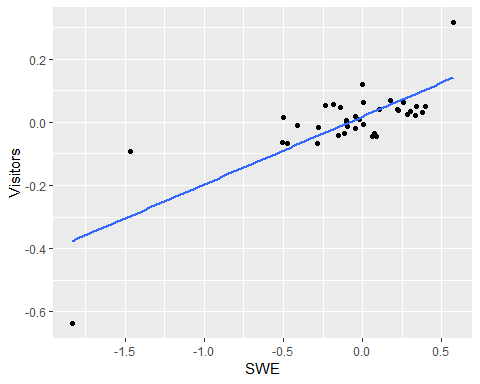


## ts.SWE.res ts.visit.res   
## Min. :-93.810 Min. :-48.7127   
## 1st Qu.:-26.661 1st Qu.: -4.5882   
## Median : -8.144 Median : -0.2427   
## Mean : 0.000 Mean : 0.0000   
## 3rd Qu.: 30.094 3rd Qu.: 7.1989   
## Max. : 62.831 Max. : 19.7768

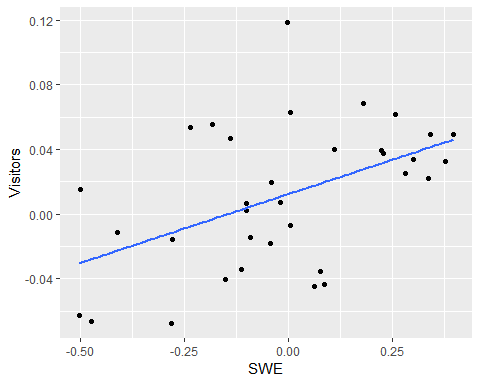
Figure 7

As shown in figure 7, there is a strong positive coorelation between the residual of the average number of ski visitors and the residual of the average amount of SWE in the same year. This shows that on years where there is lower than average snowfall, there is also lower than average ski visitors.

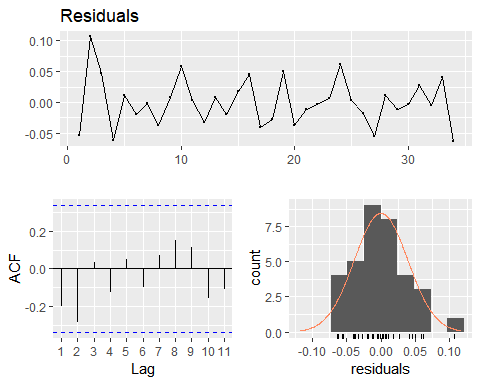
Because the visitors will continulously be going up, and there are many other factors that influence why more people seem to be skiing each year, I plotted the annual change in visitors and compared it to the annual change in SWE to compare if lower than average snowfall results in lower that average visitors. This should give a good indication whether lower than average snowfall also generates lower than average visitors to an area.



Because of the outliers adding messyness to the data, I removed them to make only the clump in the middle.



##   
## Call:  
## lm(formula = Visitors ~ SWE, data = delta.rm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.063090 -0.025171 -0.002354 0.016609 0.106532   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.012309 0.006789 1.813 0.07920 .   
## SWE 0.084925 0.026893 3.158 0.00346 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.03954 on 32 degrees of freedom  
## Multiple R-squared: 0.2376, Adjusted R-squared: 0.2138   
## F-statistic: 9.972 on 1 and 32 DF, p-value: 0.003457

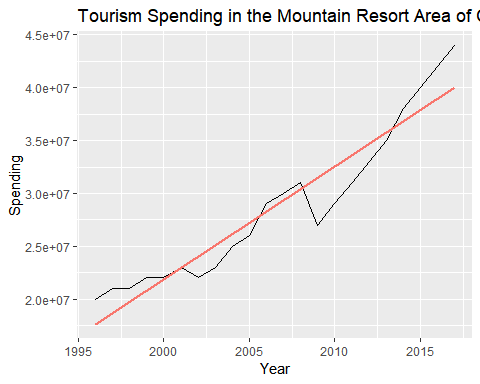


##   
## Breusch-Godfrey test for serial correlation of order up to 6  
##   
## data: Residuals  
## LM test = 13.16, df = 6, p-value = 0.04056

These charts show the average annual change of SWE and Visitor data. This shows that there is a positive coorelation between the change in snowfall from the previous year and the amount of people who visit ski areas.

# Tourism

Tourism data was manually scraped from the Colorado department of tourism, and the tourism numbers represent total tourism dollars spent in the Mountain Reort area of Colorado. Unlike the rest of the data, whcih started in 1980, this data was only available since 1996, therefore this is a smaller data set and does not contain as many years worth of data.

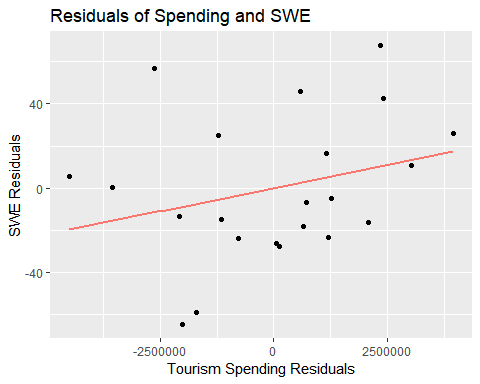
 Figue 8

As shown in the plot, like visitor data, spending in the ski resort areas of Colorado also increased ove time.however there does seem to be annual shifts from the average.

Figure 9

This plot is to show coorelations of all data points, SWE, concentration, visiter data and income from tourism since 1996. Some of these have very strong coorelations, which probably are more of a causation, such as spending and . Others however such as spending and visitors which also has a strong correlation makes sense, as more tourists would likely also mean more income from tourism.

ggplot(as.data.frame(res.all), aes(x = Spending.Residuals, y = SWE.Residuals)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE, aes(color = "red")) +  
 labs(title = "Residuals of Spending and SWE", x = "Tourism Spending Residuals", y = "SWE Residuals") +  
 theme(legend.position = "none")

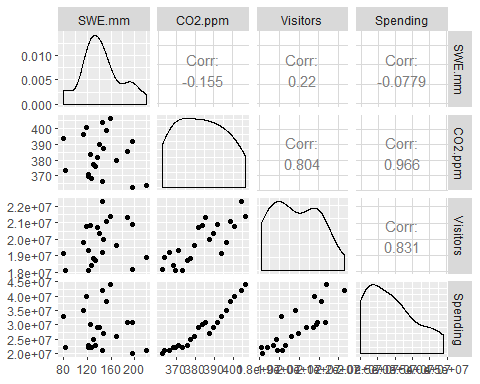


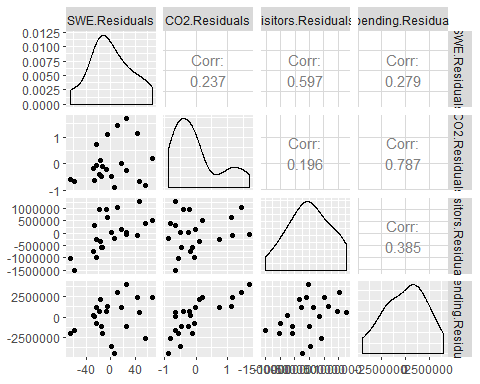
summary(lm(SWE.Residuals ~ Spending.Residuals, res.all))

##   
## Call:  
## lm(formula = SWE.Residuals ~ Spending.Residuals, data = res.all)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -55.755 -24.139 -6.885 22.886 68.295   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.081e-15 7.190e+00 0.000 1.000  
## Spending.Residuals 4.366e-06 3.364e-06 1.298 0.209  
##   
## Residual standard error: 33.73 on 20 degrees of freedom  
## Multiple R-squared: 0.07767, Adjusted R-squared: 0.03155   
## F-statistic: 1.684 on 1 and 20 DF, p-value: 0.2091

Figure 10 This plot is to show the residual data of SWE and tourism spending. As shown there is a positive coorelation between a difference from the avaerage amount of SWE and average visitor spending. Although this is the least coorelative, it does show there is a small link between the change from average SWE and the change from average tourism spending.

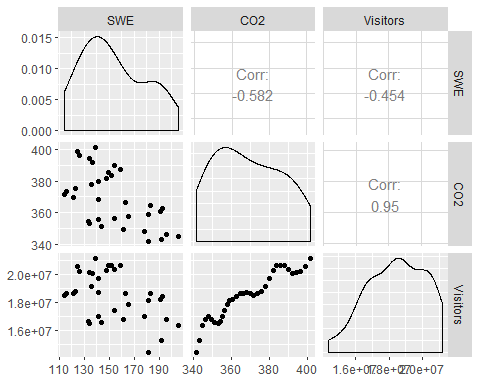
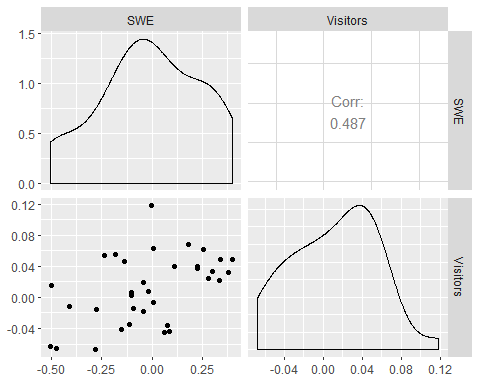
# Overview

The following plots are overviews of correlations of the data.  Figue 11

This is an overview of all data points. Because this includes data from visitors, the data included is only from 1996  Figure 12

This plot is to show the comparisons of the residuals of each data type comparatively.

From a quick look, it seems that the second best correlation is SWE and visiotrs. Also, the only histogram that is not generally in a normal distribution in this plot is residuals, which is to be expected becacuse the concentration generally is only increasing.

 This plot shows the moving averages of eveything compared.  Comparing annual change for SWE, and Visitor data.