Carbon Dioxide in the Atmosphere and Extreme Weather

Owen A. Isley
oisley@u.rochester.edu
University of Rochester
Rochester, New York

ABSTRACT

Have rising carbon dioxide levels worsened storms?

This report compares the annual mean level of carbon dioxide in the atmosphere (parts per million or p.p.m.) against the storms that took place in the United States from 1959 to 2011.

KEYWORDS

linear regression, data preprocessing, pattern recognition

ACM Reference Format:

Owen A. Isley. 2019. Carbon Dioxide in the Atmosphere and Extreme Weather. *In Proceedings of DSC 240 (Owen A. Isley)*. ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

The American government agency N.O.A.A. (National Oceanic and Atmospheric Administration) report that since sampling began (1959), the amount of carbon dioxide in the atmosphere has strictly risen.[1]

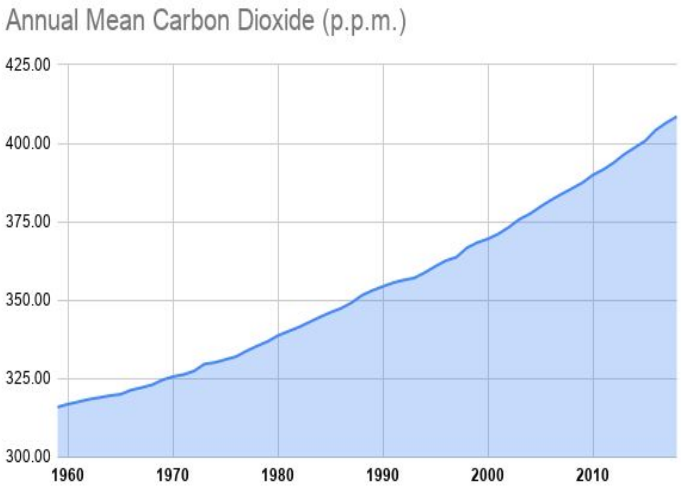


Figure 1: Mean carbon dioxide over time

As a result, scientists across the globe began to propose potential consequences of the rising carbon dioxide. Some highly publicized results include the warming of global temperatures, the widespread disruption of photosynthesis, and the worsening of storms.[2]

N.O.A.A. have also recorded every storm in the United States since 1950 in their public dataset *NOAA Storm Events Database*.[3] Using the two datasets, this report compares each year's mean carbon level with the frequency and severity of storms that year.

The objective of this report is to examine whether or not there are any patterns between rising carbon dioxide and storms in the United States.

The dataset used for analysis contains 53 records for the years 1959 to 2011 with the attributes: `YEAR`, `MEAN CARBON DIOXIDE`, `NUMBER OF STORMS`, `DAMAGE` (in dollars), and `INJURIES AND DEATHS`.

The approach was as follows:

- *Exploratory Data Analysis*: Present the composition of the dataset and the distribution of the variables
- *Data Preprocessing*: Normalize inconsistent naming/formats, calculate the number of storms per year/the total damage caused by storms/the total number of deaths and injuries, convert the dataset to a categorical (transaction-like) version for association analysis, manually classify "bad" years for k-NN classification.
- *Analysis and Modeling*: The techniques used include linear and polynomial regression, association analysis via apriori, and a customized k-Nearest Neighbors classifier.

This report will detail each step of the aforementioned process and examine the final results.

2 EXPLORATORY DATA ANALYSIS

To get an understanding of the composition of our dataset, several basic analyses were performed using the Python pandas and numpy libraries.

The exploratory data analysis consists of:

- *Data Exploration*
- *Distribution of Numeric Features*
- *Distribution of Categorical Features*

2.1 Data Exploration

2.1.1 As Numeric. Linear and polynomial regression require all the features to be numbers. The following confirms this:

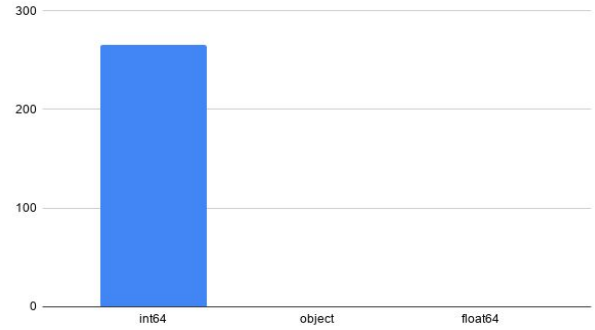
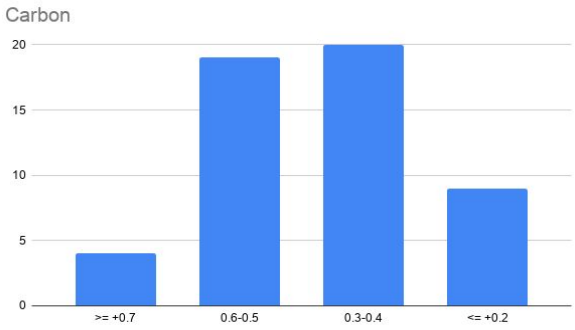
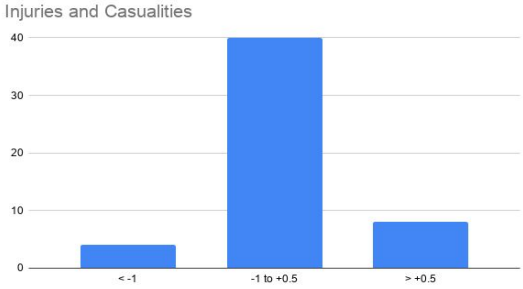
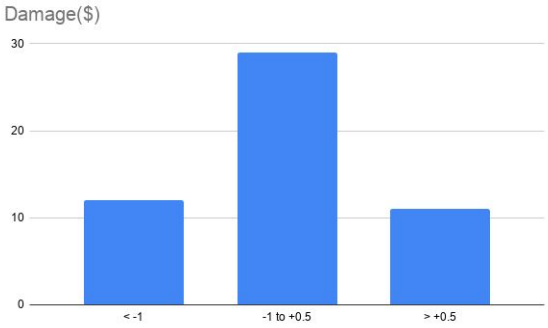


Figure 2: Features by Data Type (for regression)

2.1.2 As categorical. Association analysis requires the dataset to resemble a set of transactions. So, the numeric values were rounded to force them into "buckets:"







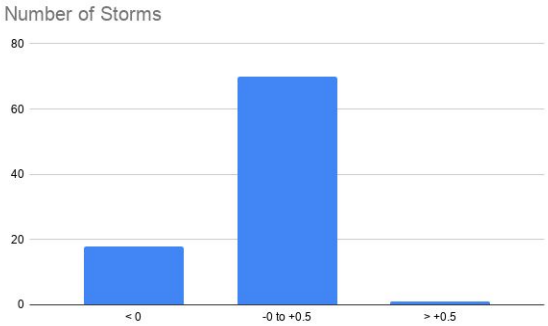


Figure 3: Features as Percent Change from Previous Year (as categories for association)

2.1.3 With class labels. Supervised k-NN clustering requires each row to be labeled for its class. The following illustrates the distribution of stormy years versus non-stormy years:

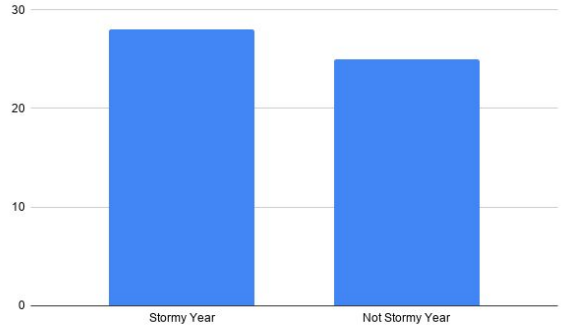


Figure 4: Features Classified (for k-NN)

2.2 Distribution of Numeric Features

The following compares the numeric features of the dataset: year, number of storms, mean carbon dioxide levels, amount of damage, and number of deaths and injuries:

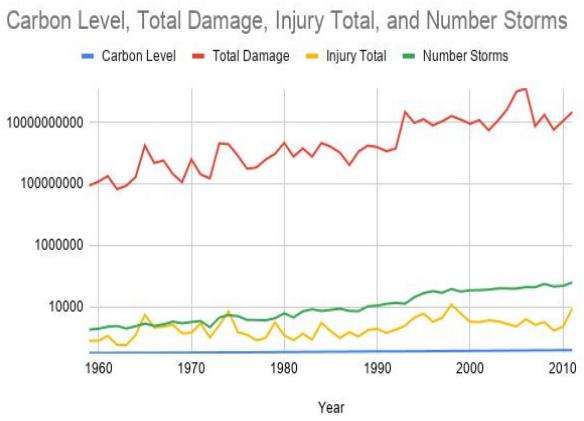


Figure 5: Distribution of Numeric Features

2.3 Distribution of Categorical Features

See Figure 3

3 DATA CLEANING, PREPROCESSING, AND FEATURE ENGINEERING

For this report, the data is examined on an annual level. Because the Storm dataset contains a row for each storm (902,297 records), it must be condensed to the annual level. This involves calculating the number of storms per year, calculating the total amount of damage (by year), and calculating the total number of injuries and casualties (by year). Additionally, inconsistently named or inconsistently formatted values were normalized.

3.1 Data Cleaning

3.1.1. .txt to .csv. The annual mean levels of carbon dioxide dataset was originally in .txt format. In order to be compatible for merging with the Storms dataset, the contents of co2_annmean_mlo.txt were read line-by-line, parsed, and exported as co2_annmean_mlo_preprocessed.csv.

3.1.2. Removal of extraneous columns. Both datasets contain a number of relevant columns. To reduce the dimensions of the datasets, only the following columns were kept from the carbon dataset: `YEAR, CARBON`; and the following from the Storms dataset: `BGN_DATE, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP`

3.2 Data Preprocessing and Feature Engineering

3.2.1 Date Conversion. In order to merge the Storms dataset with the Carbon one, they must contain a shared column (`YEAR`). So the MM/DD/YYYY format in the Storms dataset is truncated to YYYY.

3.2.2 Damage and Injuries: Conversion and Condensing. In the Storms dataset, the coefficient of the amount of damage (e.g. 1.2) is reported in one column and its magnitude in the adjacent column (e.g. 10^100000). The columns must be multiplied to determine the damage value e.g. $12,0000 in damage. Then the property and crop damage totals are added together to get a total damage for that storm.

The injuries and casualties columns are added together to get a count of the total number of injuries and casualties per storm.

3.2.3. Condensing into year-level. All rows with the same YYYY value for `YEAR` are added together to calculate the total number of injuries, the total number of casualties, the total amount of crop/property damage, and the count of the number of storms per year. The result is a dataset that has one entry per year and all of the sum totals for that year's features: `NUMBER OF STORMS, DAMAGE ($), CASUALTIES/INJURIES`

3.2.3 Merging the datasets. The Storm dataset of 902,297 records was merged with annual mean carbon dioxide dataset. Since the Storms dataset covers a smaller range of years, a left merge was performed on the Storms dataset with the Carbon dataset on `YEAR`. The result is a dataset from 1959 to 2011 that contains the attributes: `YEAR`, `MEAN CARBON DIOXIDE`, `NUMBER OF STORMS`, `DAMAGE` (in dollars), and `INJURIES AND DEATHS`.

3.2.4. Convert to categories. For the association analysis, the dataset needs to contain categorical data that resembles a set of transaction data. So, all of the

values were rounded until they began to naturally form “buckets” and then the numbers were converted into text.

3.2.5. Classify manually. For the k-NN classifier, another column was added called `CLASS` which reported 1 if the year was stormy or 0 if the year was not stormy. I decided that if a year had above the average for any of the following: number of storms, damage from storms, or injuries from storms then the year is to be classified as a bad year.

4 ANALYSIS AND MODELING

4.1 Baseline Model: Linear Regression

A baseline linear regression was performed using sklearn’s `LinearRegression` library on the following pairs:

- carbon level versus frequency of storms
- carbon level versus number of injuries/casualties
- Carbon level versus amount of damage ($)

Since there are only 52 records (1959-2011) it is necessary to train with as many records as possible. So for each regression, 10% of the dataset was used for testing and 90% for training by passing `test_size=0.1`into `train_test_split()`.

The line of best fit is compared against the six withheld records (test set). The regressor’s coefficient, the mean squared error, and the coefficient of determination was calculated for each regression.

```
Coefficient:
 [[719.05240795]]
Mean squared error: 40689424.75
Coefficient of determination: 0.74
```

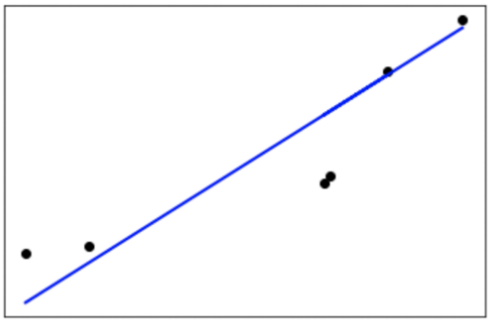


Figure 6: Regression Line (Carbon v.s. Number of Storms) with Test Set

```
Coefficient:
 [[38.2646192]]
Mean squared error: 903006.83
Coefficient of determination: 0.42
```

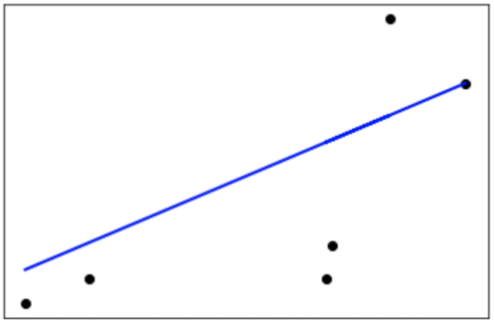


Figure 7: Regression Line (Carbon v.s. Number of Injuries/Casualties) with Test Set

```
Coefficient:
 [[5.01374283e+08]]
Mean squared error: 111792004321444052992.00
Coefficient of determination: -6.80
```

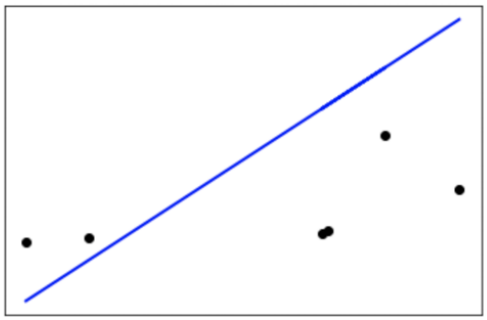


Figure 8: Regression Line (Carbon v.s. Amount of Damage($)) with Test Set

The results suggest that there is a very strong correlation between a year’s carbon dioxide level and the number of storms (coefficient of determination 0.74), a strong correlation between carbon and injuries (c.o.e. of 0.42), and no correlation between carbon and damage (c.o.e. of -6.80)!

While the linear regression did a decent job comparing carbon and frequency/injuries, from its very negative coefficient of determination, it is clear the linear regression was not appropriate for comparing carbon and damage. For this reason a polynomial regression was performed on the same three paris.

4.2 Model Enhancement: Polynomial Regression

To get a better understanding of the relationship between carbon and frequency/injuries/damage, a polynomial regression was performed on the pairs:

- carbon level versus frequency of storms
- carbon level versus number of injuries/casualties

using both sklearn's LinearRegression and PolynomialFeatures. Since there are only 52 records (1959-2011) it is necessary to train with as many records as possible. So for each regression, 10% of the dataset was used for testing and 90% for training by passing `test_size=0.1`into `train_test_split()`. It was discovered that polynomials of degree four was sufficient.

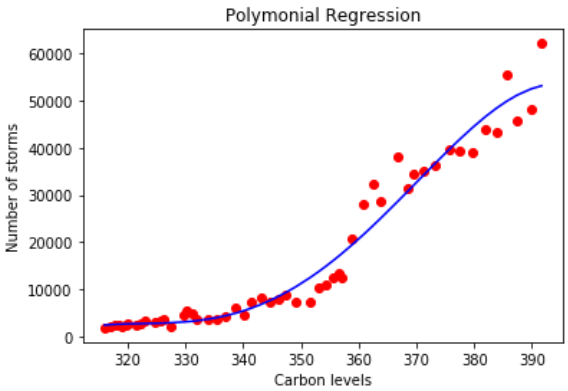


Figure 9: Poly. Regression: Carbon v.s. Number of Storms; lin_reg.score(X_train, y_train)=0.88

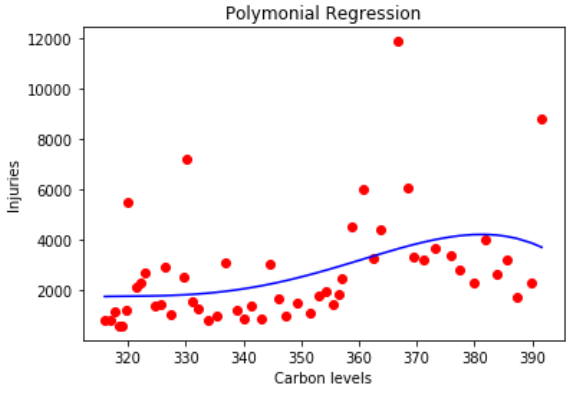


Figure 10: Poly. Regression: Carbon v.s. Number of Injuries/Casualties; lin_reg.score(X_train, y_train)=0.15

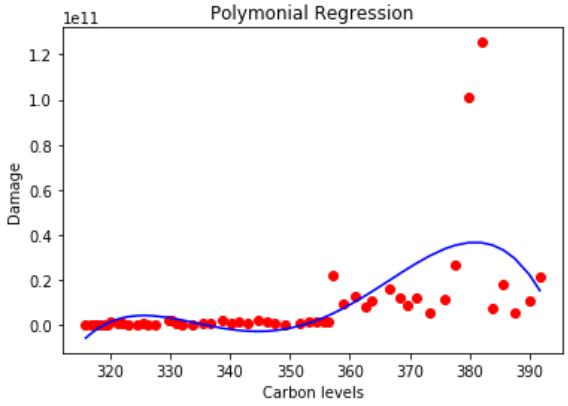


Figure 11: Poly. Regression: Carbon v.s. Amount of Damage($); lin_reg.score(X_train, y_train)=0.15

The ordinary least squares linear regression was run on the transformed dataset to evaluate the polynomial regression.

The results of the polynomial regression confirm the very strong relationship between carbon levels and number of storms, but suggest weak and very weak correlations for carbon versus injuries and carbon versus damage!

From both regression techniques, there appears to be a very strong relationship between carbon and number of storms, although a relationship between carbon and injuries/damage remains unclear.

4.3 Model Enhancement: Association Rules via Apriori

Since the regressions appear to indicate a relationship between carbon and number of storms at the large level, the next appropriate step is to examine year-by-year. Specifically to answer, *Does a percent rise in carbon dioxide in the atmosphere correlate with a specific percent rise in storm frequency/injuries/damage?*

For this technique, the YEAR column is dropped from the data set, and the values are converted into percent changes from the previous year (see Figure 3).

The percentages are rounded enough so that the values naturally bucket themselves. A letter is added to the front of each value to identify which feature that value is from. For example,

- C0.3 means ~0.3% rise in carbon from the previous year
- D30 means ~30% rise in damage amounts from storms from the previous year
- I-10 means ~10% *fall* in injuries/casualties from storms from the previous year

The numbers are converted into text and ran through the apriori Python library.

Because there are only 52 records, associating the features required a lot of rounding and a lenient minimum support, minimum confidence, minimum lift, and minimum length to generate any sort of association rules with carbon on the left side of the rule. This already indicates that any cause-effect relationship is unlikely at the year-year level.

This took an incredible amount of time to get even somewhat working.

Rule: C0.3 -> D-170.0
Support: 0.057692307692307696
Confidence: 1.0
Lift: 5.199999999999999
==========
Rule: C0.3 -> S10.0
Support: 0.09615384615384616
Confidence: 0.5
Lift: 1.625
==========
Rule: C0.5 -> D60.0
Support: 0.057692307692307696
Confidence: 0.6
Lift: 2.2285714285714286
==========

Figure 12: Some Association Rules

Even with generous rounding and lenient acceptance parameters, only nine association rules were discovered.

Many of the rules were either not relevant (did not have carbon on the left side), or contradicted themselves. For example, the top and bottom rules from Figure 12 say that a rise in carbon is associated with both a 170% drop in damage and a 60% rise in damage.

Since apriori failed to produce meaningful association rules, it is safe to suggest that a particular rise in carbon one year *is not associated with a particular rise in storm frequency/severity for that year*.

This is useful because any consequence of carbon rising is probably not realized in the year-to-year basis. While regression indicated that there is perhaps a correlation at the large-scale with frequency, apriori indicates that the correlation is not valid year-by-year.

This suggests that the impacts of rising carbon are likely not realized in the year they rise.

4.4 Model Enhancement: Custom k-Nearest Neighbors Classifier

To incorporate the long-term correlation found by the regression models with the lack of association found between the features under apriori, a k-NN classifier was implemented to see if a stormy season is predictable on its level of carbon alone.

Since the carbon levels have been strictly rising, if the classifier can assign a year as stormy or not based on just its carbon level, then it suggests that rising carbon and bad year of storms are in fact somewhat related and that we can expect the number of storms to continue to grow year-by-year.

For k-NN to train, it requires a labeling of each record as one of the classes (stormy year = 1, not stormy year = 0). To create this classification, I calculated the averages for number of storms, number of injuries, and amount of damage. If any year exceeds the average on any of the three features, then the year is classified as a 1. In total, 28 years were assigned a 1.

Since there are only 52 records (1959-2011) it is necessary to train with as many records as possible. So, 10% of the dataset was used for testing and 90% for training. It was discovered that k = 3 was sufficient.

The test records were then ran through the classifier and evaluated for accuracy. From Figure 13 we see that 3-NN classified with a “score” of 83%.

```
In [7]: knn.predict(X_test)
Out[7]: array([0, 0, 0, 0, 1, 1])

In [8]: knn.score(X_test, y_test)
Out[8]: 0.8333333333333334
```

Figure 13: Some Association Rules

5 CONCLUSIONS

The regressions show that at the large scale we see a very strong correlation between rising carbon and storm frequency. For carbon and injuries/damage, it does not appear as clear.

Therefore at the large-scale carbon dioxide appears to affect storm frequency but not storm severity.

The association rules indicate that particular percent rises in carbon does not seem to relate to that year's storms, or any particular percent rise in frequency or severity.

This is useful because it indicates that any change in carbon levels is unlikely to manifest consequences in that same year, if at all.

However, k-nn reveals that particular carbon levels alone can be used to pretty reliably classify a year as stormy or not.

Given that carbon levels have strictly risen and are strictly rising, k-NN suggests that we can expect more and more stormy years going forward.

In conclusion, it seems that if there were a relationship between carbon and storms:

1. *It appears any relationship is between carbon dioxide levels and storm frequency rather than storm severity.*
2. It appears there is not association year-to-year between carbon rising and storm frequency/severity.
3. However if a year's carbon dioxide level is high it is more likely to be a stormy year. Since the levels are strictly rising we can expect more stormy years.

6 ACKNOWLEDGMENTS

I thank Professor Pawlicki and Teaching Assistant Muhammad Ahmad for a fascinating course on the cutting edge data mining techniques. I thoroughly enjoyed this project and would recommend this course to a friend. ■

REFERENCES

[1] Dr. Pieter Tans, NOAA/ESRL and Dr. Ralph Keeling, Scripps Institution of Oceanography, "Trends in Atmospheric Carbon Dioxide." N.O.A.A., 5 Dec. 2019. co2_annmean_mlo.txt, https://www.esrl.noaa.gov/gmd/ccgg/trends/data.html

[2] Lindsey, Rebecca. "Climate Change: Atmospheric Carbon Dioxide: NOAA Climate.gov." Edited by Ed Dlugokencky, *Climate Change: Atmospheric Carbon Dioxide*, N.O.A.A., 19 Sept. 2019, www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide.

[3] National Weather Service, "Storm Events Database." N.O.A.A., Aug. 2019. repdata_data_StormData.csv, https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2